化学因子分析

潘忠孝 编著

7-312-00363-X ￥8.00

合肥 中国科学技术大学出版社 1993

目　　次

1 导　　论

1.1 计量化学的兴起

在当代科学技术革命浪潮的推动与冲击下，分析工作的仪器化、自动化和计算机化等得到迅猛的发展. 现代分析仪器能迅速、准确地为人们提供大量的可靠的量测数据，因此，对于分析化学工作者来说，如何选择最合适的分析方法和最优的测量过程，以及如何对由实验得到的原始数据进行再加工以便从中提炼出新的、更多的有价值的化学信息就显得越来越重要. 随着电子计算机科学、应用数学以及统计学方法在化学、尤其是在分析化学中应用的不断广泛和深入，一门崭新的化学分支学科 —— 计量化学 (Chemometrics) 诞生了. 计量化学的兴起，正在逐步改变以前被认为理论工作落后于迅猛发展的学科实践的分析化学的学科面貌.

计量化学是建立在多学科基础上的一门新兴学科，是化学的一个重要分支. 计量化学是在 1970 年由瑞典的 S. Wold 教授首先提出的，并与美国的 B.R. Kowalski 教授一起于 1974 年在美国 Seattle 成立了计量化学学会，他们以改进化学和应用数学及统计学衔接为目标，并把计量化学的任务规定为：应用和发展统计学方法和其它数学方法，从实验量测中取得有用的化学信息. 鉴于现代分析仪器的发展，以及计量化学在计算机与分析仪器相结合时所显示的优越性，Kowalski 进一步认为，计量化学作为一门新的化学学科，应强调应用现代数学和统计学去改进分析仪器，并从化学量测中取得更有用的化学信息. 在华盛顿成立的计量化学学会上，计量化学的初次定义为："计量化学是一门化学分支学科，它应用数学和统计学方法 (借助计算机技术)，设计和选择最优的测量程序和实验方法，并且

通过解释化学数据而获得最大限度的信息．在分析化学领域中，计量化学是应用数学和统计学方法，用最佳方式获取关于物质系统的有关信息．”如果把现代分析仪器看作是体现现代分析化学功能强弱的硬件的话，那么，计量化学就可比作是反映现代分析化学水平高低的软件．强功能的硬件和高水平的软件的结合及开发和应用，将使现代分析化学的面目焕然一新．

计量化学的发展，不但开拓了对统计学的运用，而且已涉及到光谱和波形分析、运筹学及控制论等计算机应用技术．计量化学所包括的内容相当广泛，据目前较流行的资料概括，主要有：①统计学方法、②最优化方法、③信号处理、④因子分析、⑤曲线分辨、⑥校正、⑦模型化与参数估计、⑧结构与性能相关、数据库及其检索、⑨模式识别 (包括图象分析) 和⑩人工智能等．

计量化学在解决复杂的化学问题中显示出强大的生命力，并导致了分析化学的又一次革命，引起了化学工作者的极大关注和浓厚的兴趣．开展计量化学的研究，可从实验量测数据中尽可能多地提取有用信息，实现分析工作者由过去的单纯的“数据提供者”到“问题的解决者”的飞跃．计量化学在分析化学中已有多方面的应用，它能提高分析测试的精密度与准确度．在分析化学中，无论是选用已建立的方法，还是开发新的分析方法，都要进行一系列条件实验，研究各种因素的影响，以确定最佳的测试条件，获得最佳的测试结果．在各因素间存在协同影响时，单因素实验将不能获得最佳测试条件，而这时计量化学方法却能提供有用的帮助，能借助信号处理技术提高分析测试的灵敏度与选择性，提高信噪比，消除干扰，使重叠信号得以分辨．如傅里叶变换及各种滤波技术的应用，使分析测试发生了深刻的变革．计量化学能促进仪器联机与自动化及智能化，由于电子计算机技术的发展，特别是微型计算机的普及，分析方法与分析实验室的自动化程度越来越高，采用有效的计量化学手段可使数据的获取、处理及由分析数据加工成有用的分析信息的过程日趋自动化与智能化．计量化学还可帮助化学家发展许多新的测

量方法.

计量化学是一门方兴未艾的学科，其基本理论和应用的研究仍在继续深入，很多新的原理、方法和技术仍在不断地引入这个领域中．计量化学从诞生至今仅 20 年，但其研究方法已逐渐为广大分析化学工作者所掌握和应用，国内外学者在这一领域做了大量的理论和应用研究工作，并取得了令人瞩目的进展．计量化学正在改变着以前被认为缺乏理论基础的分析化学学科的面貌，它的兴起正为广大分析工作者提供了崭新的分析方法和手段，显示出强大的生命力．计量化学已成为现代分析化学的重要的理论基础和必不可少的部分，积极开展计量化学的理论及应用研究，必将有效地促进分析化学的迅速发展．

1.2 因子分析在计量化学中的地位

因子分析是一种多元统计分析方法，最早被用于进行心理学研究，后来逐渐被引入自然科学领域．本世纪 60 年代起，因子分析被用于研究化学中的多变量问题，并已成为计量化学的最强有力的技术之一．

因子分析是通过对一数据矩阵进行特征分析、旋转变换等操作，以获得许多有关信息的数学方法．因子分析在被用于不同的学科、领域时，形成了具有各门学科特色的因子分析方法．在运用因子分析技术来研究和解决化学问题的长期实践中，经化学家及合作者的不断丰富和发展，已逐步形成了带有浓厚的化学特色的因子分析方法．在化学中，因子分析的用途可概括为：①确定影响一特定的数据矩阵的因子数，即研究和分析复杂的或是数量庞大的量测数据，确定影响这些数据的因子数；②获得对量测数据的定性的或定量的解释．

由于因子分析是一种多元统计分析方法，因此在解决多变量问题时，具有显著的优点．首先，因子分析通过对数据进行解析，可得

到影响数据的因子数目．另外，通过将变量进行组合．归结为具有实际意义的少数变量，并研究其特性与作用．通过适当的变换，可用有物理意义的参数来表示因子的本质．因子分析被引入化学中，可使化学工作者能够处理过去许多无法解决的复杂的多变量问题．概括起来说，因子分析主要有以下几个优点：①可用于很复杂的问题．因子分析作为一种多变量分析方法，可同时能处理许多因素相互影响的复杂体系．这一特点在化学中特别重要，因为对大多数化学数据的解释要借助于多变量手段．②能快速地对大量数据进行处理．借助电子计算机，使用标准的因子分析程序，能快速地分析大批量数据．③能研究多种类型问题．不论是在对原始分析数据了解甚少或是在对数据的本质一无所知的情况下，都可应用因子分析方法．结合理论学说来应用因子分析自然是理想的，不过，以经验与直觉为基础，因子分析技术也可产生有价值的预测．④可压缩数据，提高数据质量．通过对数据矩阵进行因子分析，可用最少的因子来表示它们，而基本上不损失数据原来所包含的信息，并且还可发掘出某些潜藏的规则．⑤可获得对数据的有意义的解释．通过因子分析可对样品或变量进行分类；能够为体系建立完整的有物理意义的模型，可预测新的数据点；通过目标检验，还可进一步作定量研究．总之，对于那些看起来很难解决的复杂化学问题，因子分析技术往往可给我们提供有益的帮助．

因子分析在化学中的应用相当广泛．通过处理色谱、质谱、核磁、红外、紫外－可见、喇曼、萤光等光谱数据以及 X 射线光电子能谱和俄歇电子能谱数据，可以对待测体系进行定性定量分析，也可加深对这些谱学本身的基础理论研究．此外，又可用于研究化学平衡及化学动力学等问题．

因子分析是计量化学中一种十分有用的多元统计分析方法，除了用于解决一些多元统计问题外，还可用于计量化学中的其它许多领域，如曲线分辨、校正、模式识别等．熟悉和了解因子分析的基本原理对于理解和更好地应用其它计量化学方法是有益的．

因子分析作为一种多元统计分析方法，其理论至今尚待完善，因此国外化学界对因子分析的理论模型及其在化学中的应用研究开展得颇为活跃．相比之下，在我国，化学中的因子分析理论及应用研究（尤其是应用的广泛性）一则起步较晚，再则是涉及的面也较狭窄，使得这种已被实践证明是解决化学中较复杂问题的有力的计量化学工具没有得到广泛的掌握和运用．因此，结合我国的现有条件，积极广泛地开展化学因子分析的理论及应用研究对发展和促进我国化学、尤其是分析化学学科的发展是具有重要意义的．

1.3 化学因子分析理论的发展

因子分析的模型最早由 J. Pearson 和 C. Spearman 提出，首先应用于心理学研究．由于这种研究收到了较好的效果，因而引起了科学界的注意．数十年来许多统计学家以及其它科学工作者在因子分析的理论、方法和实际应用等方面做了大量的工作，使因子分析不断得到充实并成为多元统计学的重要组成部分．与此同时，因子分析的应用也逐渐推广到心理学以外的其它学科，如经济学、生物学、植物学、地质学等．化学家对因子分析的应用和研究始于本世纪 50 年代末和 60 年代初．

R.M. Wallace 首次用矩阵求秩法研究混合物的吸收光谱数据以确定体系中存在的吸光物种数．矩阵求秩法的基本思想是通过矩阵代数和统计规则或误差判据求出吸光度数据矩阵的秩数，其秩数即为体系中存在的吸光物种数．矩阵求秩法通过得到矩阵秩数进而获得体系中共存的吸光物种数的信息．显然，对于较大的矩阵，其求解过程显得十分麻烦，同时该法未能提供秩之外的任何其它信息．因此，它的应用受到了限制．然而求秩法毕竟是最早被用于确定吸光物种数的一种因子分析技术，因此，应当认为，这一工作奠定了因子分析在化学中的应用基础．

受前人的启发，J.J. Kankare 在 1970 年报导了他首次用抽象因

子分析法 (AFA) 处理配位化合物的吸光度数据以确定体系中存在的吸光物种数. 随后，抽象因子分析法在化学中的其它领域 (如色谱、质谱、核磁等) 也得到应用. 抽象因子分析不仅能确定矩阵的秩数，而且还可获得对数据的定性解释，可用于对样品或变量进行分类等，是其它因子分析技术的基础.

在因子分析过程中，因子数的准确确定是极其重要的. 由于分析数据 (通常由实验测得) 存在误差，这就给确定因子数带来很大困难. 因此，如何在掺和了误差的实验数据中准确地找出影响原始分析数据的因子数，是因子分析研究中的一个特别重要也很困难的研究课题. 多年来，不少致力于因子分析理论研究的化学工作者在这方面作了很大的努力. 然而，迄今，还没有现成的准确确定一套实验数据中有多少有意义的因子的严格的方法. 不过，已有不少判据在确定因子数时是行之有效的，其中被应用得较多的有 E.R. Malinowki 等提出的判据 (如 IE, RE, IND 等). 用交互校验和频率分布等方法确定因子数也是相当有效的. 这一方面的研究尚有待进一步的深入.

通过对原始数据矩阵进行抽象因子分析，可获得影响数据的因子数以及数据的随机误差的重要信息. 但是，抽象因子分析得到的是纯数学上的结果，缺乏明确的物理或化学意义，若需进一步了解影响数据的各因子的本质，则需要对抽象的因子进行旋转变换，使其具有明确的物理或化学意义. E.R. Malinowki 等提出了一种称为目标因子分析的方法 (TFA)，该法通过目标检验，将抽象的因子转换为实际因子. 目标因子分析是解决化学问题的一种颇为适用的方法，它不但适用于定性分析，而且还适用于定量分析，已在化学研究中得到广泛的应用. 例如，与光度法结合可进行多组分混合物的同时测定等.

一般的因子分析方法是以原始分析数据是各因子的线性加和这一假设为基础的. 然而，在处理实际问题时，往往遇到变量是非线性关系的情况，为了克服这种线性假设的局限性， C. Jochum 和 B.R. Kowalski 建立了包括线性和非线性的因子分析程序，把它简称

UVFA. UVFA 涉及的内容相当丰富，非线性因子分析只是其中一部分，它是以非线性最小二乘投影原理为基础的，涉及两种数学方法，即多维换算和参数图方法.

现代分析仪器可提供多维量测数据，在处理双线性数据矩阵 (如激发 - 发射光谱) 方面， C.N. Ho 等人提出的秩消因子分析方法 (RAFA) 是一种新颖的解析二维数据矩阵的方法. 在不考虑分析体系中其它对分析工作者完全不感兴趣的成分的情况下，要想直接获取所感兴趣的某些组分的定性和定量信息，这时， RAFA 更能显示出明显的优越性. RAFA 法已用于分析激发 - 发射、LC/UV 等二维数据的分析. E. Sanchez 等人在 RAFA 的基础上发展的广义秩消因子分析方法 (GRAFA) 包括一般的秩消法，该法将未知样品的二维数据矩阵减去一个已知组分和含量的二维数据矩阵，然后进行较复杂的特征分析以获得未知样品的组成及含量等信息. GRAFA 的出现为解析多维数据矩阵提供了强有力的工具.

为通过解析光谱滴定数据来研究配合物平衡问题，H. Gampp 等提出一种崭新的因子分析模型 —— 渐进因子分析 (Evolving Factor Analysis，简称 EFA). EFA 已成功地用于处理光谱滴定数据和 ESR 等数据. 我们将 EFA 用于处理电位滴定数据也取得了较为满意的结果. 用 EFA 处理光谱滴定数据，不仅能确定平衡体系中存在的总的物种数，而且还能指出各种吸光物质的存在范围，这是其它方法以及经典的因子分析法所无法做到的. 通过进一步迭代分析 (或变换) 还可给出各物种的光谱及浓度. EFA 在处理滴定数据方面有着很广的应用前景.

由法国的 Jean-Paul Benzecri 在 60 年代初发展起来的一种几何技术，称为对应因子分析，可被用来描述变量之间、样品之间以及变量和样品之间的双重关系. 由于数据的预处理对于样品和变量呈对称性，因而使得样品和变量可以同时被描绘在同一个二维图上. 实践证明，这种技术在化学基础理论及应用研究中有独特的用途.

综上所述，经过化学家及其合作者近 30 年的共同努力，因子

分析技术在化学研究的实践中得到了很大的丰富和发展，许多因子分析技术带有浓厚的化学特色，已基本形成了一套化学因子分析方法．随着化学工作者对因子分析技术认识的逐步加深和已有的因子分析方法应用的不断深入，适合於化学研究的因子分析理论将不断得到完善．

2 抽象因子分析

随着研究的进展，结合不同学科的实践，已发展了多种因子分析方法．考虑到因子分析法发展的历史及为了阐述某些在各种不同方法中均是共通的基本步骤，我们首先介绍抽象因子分析法．随着讨论的深入，便不难明白所谓“抽象”的意义．

为了阐述上的方便及合理性，在本章中，有时也会涉及其他因子分析方法的部分内容．

2.1 因子分析的主要步骤

2.1.1 主要步骤概述

概括地讲，因子分析包括三大步骤：预备、复原和变换．图 2.1 给出这几大步骤的顺序及从每一步骤的结果所获得的有关信息．

在预备步骤中，对要进行因子分析的数据仔细地做数学上的选择和预处理．可以毫不夸张地说，仔细的数据选择是成功的因子分析的一个关键前提．在复原步骤中，作为因子分析的基础，需计算主因子解和确定因子的正确数目．由于紧接在后面的所有种种因子分析步骤都以这一步所得到的模型为基础，故复原步骤须格外小心地去进行．如果所解决的问题只要求获取的信息是因子的数目时，预备与复原步骤便构成完整的因子分析流程．然而，从复原步骤所得到的抽象形式下的行和列因子并不认为是物理的或是化学的参数，因为它们并不具有明确的物理的或化学的意义，而仅仅是纯粹的数学解．在大多数的科学问题中，主要的目的是想对因子的性质有更深一步的了解，因此，就必须设法将抽象的解变换成更有意义的解．自然，这也就构成了变换步骤的主要内容．

用相应的变换手段. 变换技术包括抽象旋转和目标变换两大类，当然，经变换后的因子也能复原原始数据.

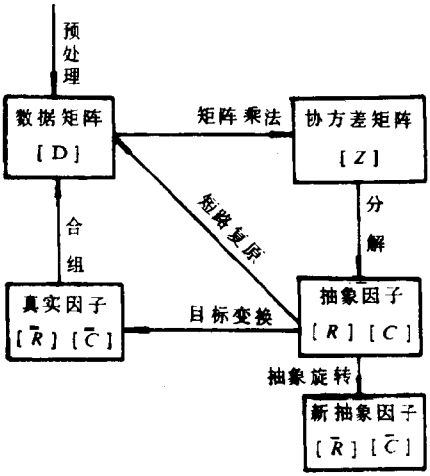


图 2.2　因子分析的主要操作示意图

本书中，$D_i' = [d_{i1}\ d_{i2} \cdots d_{ic}]$ 代表矩阵 $[D]$ 的第 i 个行向量.

2.1.2　主要步骤的数学提要

对因子分析的主要步骤有一个大概了解之后，现在介绍一下这些步骤中所涉及的较具体一些的数学要点. 至于详细的阐述或必要时所作的推导将分别在另外相应的章节中进行.

协方差矩阵 $[Z]$ 可由原始数据阵的转置左乘原始数据阵来构造

$$[Z] = [D]^T[D]. \tag{2.1}$$

找到一个矩阵 $[Q]$, 使 $[Z]$ 对角化，这样一来

$$[Q]^{-1}[Z][Q] = [\lambda_j \delta_{jk}] = [\lambda], \tag{2.2}$$

由此可得出结论：能对角化协方差矩阵的矩阵的转置就是我们所指的列矩阵．由于这一矩阵 $[Q]$ 的每一个行是一个特征向量，故通常被叫做特征向量矩阵．又因为特征向量是正交的，故

$$[C]^{-1} = [C]^T. \tag{2.12}$$

根据式 (2.6) 和 (2.11)，我们可计算行矩阵 $[R]$

$$[R] = [D][Q]. \tag{2.13}$$

得到 $[R]$ 和 $[C]$ 之后，我们便可复原数据矩阵．到此，图 2.1 所示的短路复原便已在抽象的意义下基本完成．

因子分析的一个初衷是要用最少的特征向量在误差范围内复原数据．事实上，由于实验误差的不可避免，对协方差矩阵 $[Z]$ 分解的结果总是产生与 $[Z]$ 的阶数相同数目的特征向量和特征值．在以后的讨论中我们将会指出：特征值的大小是对其相应的特征向量的重要性的度量．与最大的特征值相对应的特征向量是最重要的，而与最小的特征值相对应的特征向量是最不重要的．要在实验误差范围内复原数据，通常并不需要矩阵 $[Q]$ 的所有的特征向量，而应该剔除那些最不重要的特征向量，因为它们的保留只会产生实验误差．

2.2 预　　备

预备步骤的目的很简单：为了得到一个最适合于因子分析的数据矩阵．这一步骤的主要内容包括所要解决的问题的选择，数据的选择和对数据进行数学预处理以确认其对理论的和统计学的判据的合适性．因子分析的最终成败很大程度上取决于预处理的步骤．

2.2.1 问题的选择

在这里只打算讨论问题选择的一些总的方面．在进行研究之前，首先要从因子分析的角度对问题做仔细的评估，尽量减小分析工作起始阶段的盲目性．

1. 适合因子分析的可能性

首先要注意的是数据是否可由项积线性加和来进行模型化. 只要是在一个量测可被表示成项积的线性加和的情形，因子分析都可被应用. 是否决定应用因子分析，在可能的情况下应以理论概念为基础，如果缺乏理论概念，则应以一个对于数据来说是合理的概念模型为基础. 要是研究人员对问题有一预先的直觉，则因子分析的结果便较易于评价. 鉴于不能事先知道对某一问题是否可应用因子分析去解决，因此，因子分析过程往往只好在带有一定盲目性的情形下开始. 不过，这一点并不十分要紧，研究人员在执行复原步骤之后便能更好地判断分析工作是否应该终止或继续进行.

在众多类型可能具有因子分析解的数据中，有两种特别适合于因子分析. 包含有几种多组分混合物在几个波长处的光谱强度的矩阵往往都具有因子分析解，这是第一种. 大家都知道，在稀溶液状态下，吸光物种的浓度与吸光度之间的关系是符合比耳定律的，即每一个吸光度数据都可被表达成与因子分析模型相一致的项积的加和. 第二种是溶质 - 溶剂问题. 对于这类问题，如果怀疑所观察到的现象产生于包含有溶质 - 溶剂交互作用项在内的化学交互作用的加和的话，便可直观地期待其有因子分析解. 将溶质余因子和溶剂余因子与各单个的分子的特性相关联，构成与因子分析相一致的模型，便可解释大量的化学数据.

2. 问题的类型

许多类型的数据可用矩阵来表示. 适合于因子分析的问题可根据以数据阵的行标和列标表示的指定类型来加以分类. 有 3 种类型的指定 (物质的实体、物质的性质和时间的过程) 可被用来描述化学现象. 实体一词包含从亚原子颗粒至星系的物质的任何样品，分子、混合物、仪器和人等都是实体的例子. 性质则是实体的特征化，它可以区别实体. 光谱的间隔、色谱的间隔和强度、温度等都是性质的例子. 过程则指定量测被进行的时刻. 基于上述的 3 种指定分类，可组成 6 种类型的数据矩阵：实体 - 实体、实体 - 性质、实体 - 过

程、性质 - 性质、性质 - 过程和过程 - 过程. 例如，在一个实体 - 性质矩阵中，一套指定与一组实体相关联，另一套则与一组性质相关联. 在化学中，实体 - 性质、实体 - 实体，有可能时也包括实体 - 过程，这 3 类矩阵是最重要的. 在行为科学中，迄今研究得最多的是实体 - 性质矩阵. 在抽象因子分析文献中，这种矩阵当实体是行指定时被称为 R 型矩阵，在性质是行指定时被称为 Q 型矩阵.

重要的实体 - 性质和实体 - 实体矩阵的例子列于表 2.1 中.

表 2.1 重要数据矩阵类型的例子

矩阵类型		元素
实体 - 性质对	分子 - 物理性质	物理性质的值
	分子 - 光谱间隔	光谱强度
	溶液 - 光谱间隔	光谱强度
	混合物 - 色谱间隔	色谱响应
实体 - 实体对	溶质 - 溶剂	光谱强度
	溶质 - 溶剂	色谱保留
	样品 - 化学元素	分析的浓度

在表 2.1 中，左边给出行指定 / 列指定对的性质，右边给出供因子分析的对应数据类型. 分光光度法测定所得的吸光度矩阵是一个包含有各种混合物 (实体) 在几个波数 (性质) 的实体 - 性质矩阵. 一个包含有两种实体 (溶质和溶剂) 的色谱数据矩阵是一个实体 - 实体矩阵. 许多其他有用的数据矩阵的例子将在以后的各个章节中予以讨论.

从表 2.1 中可以见到，混合物的光谱和色谱可被用来形成实体 - 性质矩阵. 分析化学家对这类实体 - 性质矩阵更感兴趣，因为混合物中的组分数目和组分本身都可从因子分析中获得. 实体 - 实体矩阵，虽然不普遍，但物理化学家却特别感兴趣，因为因子分析能提供对实体 - 实体的相互作用的深入了解. 实体 - 过程矩阵的因子分析能对化学动力学中的有关问题提供新的解决途径.

3. 要求获取的信息

7.1.2 节)、但这并不能取代可靠的数据. 除非化学家对误差有一个合理的估计, 否则, 对因子分析的结果便不能进行令人信服的解释. 当数据是在几个实验室中被采集时, 应仔细挑选不自相矛盾的数据, 删去模棱两可的数据. 独特性检验 (见 3.6 节) 手续对检测那些有严重误差的点是有用的.

应该组合完整的数据矩阵. 缺乏足够的数据往往妨碍化学工作者去执行因子分析. 需注意, 在一个矩阵中虽然有某些点可能丢失, 不过只要能形成一个较小的完整子阵, 则因子分析仍然可以进行. 若能仔细地挑选子阵中被表示的指定, 仍可获得成功的因子分析解. 对于从统计学判据估算所得的数据点应避免采用.

数据矩阵的大小取决于数据的来源、研究的目的和手头所拥有的计算工具. 为保证能获得全面的解, 也许会采用可能得到的最大的矩阵. 不过, 过大的矩阵在开始时可能会因为过于复杂而无法分析. 为了取得某些进展, 可以有目的地用一挑选过的子阵来工作. 这样, 连续地分析大的矩阵便可获得较全面的解. 如果受到计算工具的局限, 那么, 只好将矩阵减小至可以运算的程度. 保留在子阵中的指定应该是原来的指定集的有化学意义的代表. 当考虑到要进行目标检验时, 则行和列指定的数目最好都应至少是因子数目的两倍. 这一经验规则帮助我们保证目标检验的数学正确性.

矩阵中的指定应该充分地代表化学工作者希望研究的问题. 如果许多分子都有数据可供采用, 则所选择的分子应该复盖研究者感兴趣的性质范围. 包罗太多类型的指定可能会把问题不必要地弄得太复杂. 例如, 某些具有异常化学性质的分子会给因子分析过程带入多余的因子. 除非研究人员对这类分子特别感兴趣, 否则, 这类分子是不应该被引入矩阵中去的.

2.2.3 数据的预处理

数据的预处理应尽可能以合理的理论判据为基础. 不严密的预处理会严重地损坏对结果的解释, 甚至可能会使因子分析变得无效.

一般地说，数据预处理包含以下几个方面.

1. 协方差矩阵和相关矩阵的选择

这方面存在着 4 种不同的途径可供讨论：对于原点的协方差，对于平均值的协方差，对于原点的相关和对于平均值的相关. 它们之间的联系取决于一个简单的线性变换

$$[D]^{\#} = [D][A] + [B], \tag{2.14}$$

式中，$[D]^{\#}$ 代表被处理过的数据，它们被用来进行因子分析. 上述 4 种不同的途径在 $[A]$ 和 $[B]$ 的定义有所不同. $[A]$ 是用来调节每一数据列的总的数量级的一个对角矩阵，它仅由对角元素 a_{jj} 组成. $[B]$ 是一个在任何一列中元素 b_{ij} 都相等的矩阵. 它移动因子空间的原点. 对应于这 4 种不同的途径，$[A]$ 和 $[B]$ 的定义分别为：

对于原点的协方差

$$\left.\begin{aligned} a_{jj} &= 1, \\ b_{ij} &= 0. \end{aligned}\right\} \tag{2.15}$$

对于平均值的协方差

$$\left.\begin{aligned} a_{jj} &= 1, \\ b_{ij} &= -\hat{d}_j. \end{aligned}\right\} \tag{2.16}$$

对于原点的相关

$$\left.\begin{aligned} a_{jj} &= (\sum_{i=1}^{r} d_{ij}^2)^{-1/2}, \\ b_{ij} &= 0. \end{aligned}\right\} \tag{2.17}$$

对于平均值的相关

$$\left.\begin{aligned} a_{jj} &= [\sum_{i=1}^{r} (d_{ij} - \hat{d}_j)^2]^{-1/2}, \\ b_{ij} &= -\hat{d}_j a_{jj}. \end{aligned}\right\} \tag{2.18}$$

式中，$\bar{d}_j$ 是原始数据矩阵第 j 列实验数据点的平均值，r 是一列数据中的点的个数，d_{ij} 是实验数据点.

究竟是对原始数据作因子分析 (采用协方差途径) 或是在作因子分析之前先对数据作归一化处理 (采用相关途径)，要做出上述决定必须以数据中的误差类型为根据. 当每一数据列的绝对误差相似时，采用协方差途径；当每一数据列中的相对误差 (百分误差) 相似时，采用相关途径. 如果化学工作者对问题中的误差类型无法确定时，最好是采用协方差途径. 因为大多数的化学量测包含有绝对的而不是相对的误差. 只有当数据点在数量级上相似并含有一致的误差时，两种途径方可导致相同的结果.

4 种途径的优缺点已有人讨论过. 在化学问题研究中，对于原点的协方差更受欢迎，因为它保持了因子空间的原点、因子轴的相对长度和相对误差，这恰好合乎我们需要的情形. 使用对于平均值的协方差或相关，我们会丢失与实验标度的零点有关的信息. 较多数据点的添加会使因子空间的原点产生偏移. 采用相关途径 (对于原点或对于平均值)，则会丢失与各数据列有关的相对大小和相对误差的信息.

2. 平衡在矩阵中的误差

从理论上讲，所含误差全部都一致的数据矩阵是最合适于因子分析的. 如果相差数量级的误差离散在整个矩阵中，则应构成具有更加一致误差的较小的矩阵，如这样也办不到的话，就要采用 chi 平方判据 (见节 7.3.1) 来确定因子的大小. 如果矩阵中列与列之间的误差变化很大而在每一列中却相对地是恒定的话，建议采用列的“标准化”. 每一个元素被它所在列的标准偏差去除，使整个矩阵中的偏差成同一单位. 例如，有一矩阵，它的 4 个列具有完全恒定的绝对误差分别为 1, 4, 10 和 5，那么，各对应列的每一个元素应分别被 1,4, 10 和 5 去除，而且应该采用协方差途径. 当矩阵中的列涉及不同单位时，应该进行标准化.

3. 为数据选择函数形式

除非理论或经验对数据提出了最合适的函数形式，否则应对原始数据(也许是数据的对数)进行因子分析，运用其他函数形式所作的分析通常达不到什么有用的目的．在缺乏其他判据时，化学工作者应该采用这样一种函数形式，它允许将因子分析的结果同对该课题以前所做的研究加以比较．数据的对数变换可能是最有用的，因为在化学中对数关系相当普遍．某些经验方程，如 Hammett 函数，认为平衡常数和速率常数的对数而不是原来的常数本身应该被用来进行因子分析．

2.3 数据的分解与复原

抽象分解和复原步骤是因子分析的数学基础．它们的目的是计算主因子解和确定因子的正确数目．最后所得的抽象解用主因子矩阵来表示．主要步骤示意图 2.3 中．

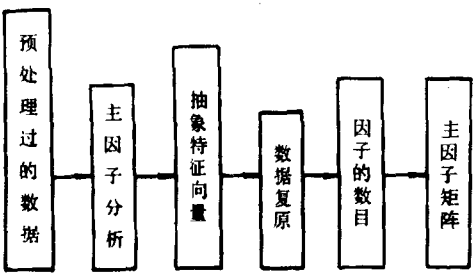


图 2.3 分解与复原流程框图

2.3.1 协方差矩阵

为了以后叙述及推演的方便，我们先在这里讨论一下有关协方差矩阵的内容．在化学问题中，协方差途径是比较重要的一种途径，尤其是对于原点的协方差途径更受化学工作者的欢迎．

数据矩阵或是它的转置是否被涉及，相同的特征向量和特征值将准确地在因子分析中出现．如果采用的是相关阵而不是协方差阵，那情形就不是这样了．

2.3.2 分解的原理

协方差矩阵或相关矩阵的秩，也就是因子空间的维数，可通过分解它们成一套特征向量而在数学上予以确定．如果数据是纯粹的，将准确的得出 n 个特征向量，n 为该数据矩阵的因子空间维数．然而，由于实验误差的存在，则所得的特征向量数目将等于矩阵的行数或列数两者之中的最小者，为讨论方便，我们以后都设数据矩阵的列数 c 小于行数 r．也就是说，有实验误差存在时，分解结果将准确的得到 c 个特征向量．为弄明白在分解步骤中所涉及的理论原理，在这里，引入基的概念．令 $U_1, U_2, \cdots, U_c$ 是单位向量基集，它旋转因子空间 [即，$U_1 = (1,0,0,\cdots,0)$，$U_2 = (0,1,0,\cdots,0)$，$U_3 = (0,0,1,0,\cdots,0)$, 等等]. 这些向量是正交的，所以

$$U_j U_k = \delta_{jk}, \tag{2.21}$$

式中，δ_{jk} 是 Kroenecker δ.

在因子空间中，一个数据点可被当作一个向量，因此

$$d_{ik} = \sum_{j=1}^{c} r_{ij} c_{jk}, \tag{2.22}$$

可以更好地表达成

$$D_{ik} = \sum_{j=1}^{c} r_{ij} U_j c_{jk}. \tag{2.23}$$

式中，d_{ik} 是该数据点.

协方差矩阵 $[Z]$ 的元素 Z_{ab} 可由下式给出

$$Z_{ab} = \sum_{i=1}^{r} D_{ia} D_{ib} = \sum_{i=1}^{r} (\sum_{j=1}^{c} r_{ij} U_j c_{ja})(\sum_{k=1}^{c} r_{ik} U_k c_{kb}), \tag{2.24}$$

因为 $U_jU_k=\delta_{jk}$，所以式 (2.24) 变成

$$Z_{ab}=\sum_{i=1}^{r}\sum_{j=1}^{c}r_{ij}^2c_{ja}c_{jb}. \tag{2.25}$$

由上式可以见到，一个完全的协方差矩阵可以被分解成列向量的加和乘以它们的对应特征值

$$\begin{aligned}[Z]=&\sum_{i=1}^{r}r_{i1}^2\begin{bmatrix}c_{11}\\c_{12}\\\cdots\\c_{1c}\end{bmatrix}[c_{11}\ c_{12}\ \cdots\ c_{1c}]\\&+\sum_{i=1}^{r}r_{i2}^2\begin{bmatrix}c_{21}\\c_{22}\\\cdots\\c_{2c}\end{bmatrix}[c_{21}\ c_{22}\ \cdots\ c_{2c}]+\cdots\\&+\sum_{i=1}^{r}r_{ic}^2\begin{bmatrix}c_{c1}\\c_{c2}\\\cdots\\c_{cc}\end{bmatrix}[c_{c1}\ c_{c2}\ \cdots\ c_{cc}].\end{aligned} \tag{2.26}$$

用简单记号表示，则式 (2.26) 可写成

$$[Z]=\lambda_1C_1C_1'+\lambda_2C_2C_2'+\cdots+\lambda_cC_cC_c', \tag{2.27}$$

式中

$$\lambda_j=\sum_{i=1}^{r}r_{ij}^2, \tag{2.28}$$

$$C_j'=[c_{j1}\ c_{j2}\ \cdots\ c_{jc}]. \tag{2.29}$$

式 (2.28) 中，λ_j 是与特征向量 C_j 对应的特征值．在此，注意到特征值是每一个行指定在其适当的特征向量上的投影平方的加和，因此，代表了单位向量的相对重要性．

因为有 c 个数据列，所以在分解过程中会出现 c 个特征向量，这也是由于数据中存在的实验误差所引起的．事实上，只有这些向量中的 n 个 (即那些与最大的特征值对应的) 被要求用来说明数据，而其余 $(c-n)$ 个特征向量是实验误差的结果．当然，数据中如果没有实验误差，则在分解过程中将会准确的出现 n 个特征向量，重要的特征向量数目代表因子空间的维数．从数学观点来讲，这意味着数据阵的真实秩等于因子空间的维数．

有许多分解协方差阵的数学方法，这里只重点介绍主因子分析和 JACOBI 这两种方法．

2.3.3 主因子分析

主因子分析法是一种最小二乘技术，在文献中常用其英文的缩写 PFA 来表示，有时也被称为主成分分析 (PCA)，它被普遍用来进行矩阵的特征分析．

主因子分析产生由一个抽象特征向量集和一个相应的特征值集所组成的一个抽象解，每一个主要的特征向量代表一个抽象因子．在因子分析中，由于向量简单地就是数的一个一维数组，故我们交替地使用“因子”和“向量”这两个名词．每一个特征值测量出其相应的特征向量的相对重要性．一个大的特征值标明出一个主要的因子，而一个非常小的特征值标明着一个不重要的因子．

正如前面已讲过的那样，如果数据不包含有实验误差，则主因子分析将准确地产生 n 个特征向量，每一个特征向量对应于一个控制因子．然而，由于实验误差的不可避免，对化学问题的主因子分析总是产生出 c 个特征向量，每一特征向量对应于数据矩阵 c 个列中的一个列．必须指出，在这 c 个特征向量中，只有与 n 个最大特征值相对应的那 n 个特征向量才具有物理或化学意义．

主因子分析可将数矩阵分解成一个抽象行矩阵和一个抽象列矩阵的积，可应用相应的因子分析计算机程序来计算完整的主因子分析解

$$[D] = [R]_{C,\mathrm{PFA}}[C]_{C,\mathrm{PFA}}, \tag{2.30}$$

由于有 c 个因子，所以，行矩阵 $[R]_{C,\mathrm{PFA}}$ 中有 c 列，列矩阵 $[C]_{C,\mathrm{PFA}}$ 中则有 c 行．这种完全的解超越真实因子空间．它包含的特征向量数目大于实际上所要求的数目．

抽象列矩阵 $[C]_{C,\mathrm{PFA}}$ 中的每一行本来就是一个特征向量，鉴于主因子分析的最小二乘性质，在重要性的降序排列中，应用特征向量就可形成列矩阵．根据因子在对数据偏差所起的作用大小可将它们分级．抽象列矩阵中的第一行 (代表第一个因子) 对应于最大的最重要的特征值．第 c 行是最不重要的，它与最小的特征值相对应．第一个因子说明数据中的方差的最大百分率，第二个次之，第三个则更次之，余者类推．这样一来，全部 c 个抽象因子准确地解释了数据，也包含了量测中的实验误差．

在主因子分析中，特征向量是被连续地进行计算的，这样可使每一步中的残余误差为最小．每一个连续的特征向量说明了在数据中的一个最大的方差．为了阐述主因子分析所涉及的哲学和数学的原理，我们在此将描述出它的全部推演过程．

为找出抽象的特征向量，需涉及到下述的数学推理．主要的特征向量构成一个最佳的、相互正交的坐标系，每一个连续的特征向量说明数据中最大的可能方差．方差的定义为

$$\text{方差} = \frac{\sum_{i=1}^{r}\sum_{k=1}^{c} d_{ik}^2(j)}{\sum_{i=1}^{r}\sum_{k=1}^{c} d_{ik}^2}, \tag{2.31}$$

式中 $d_{ik}(j)$ 是 $[D_j]$ 的一个元素，d_{ik} 是原始数据 $[D]$ 的一个元素

$$[D] = [D_1] + [D_2] + \cdots + [D_j] + \cdots + [D_c], \tag{2.32}$$

$$[D_j] = R_j C_j'. \tag{2.33}$$

这里，R_j 和 C_j' 是与特征值 λ_j 相应的行指定和列指定．

与最大最重要的特征值对应的那个特征向量在因子空间中被旋转，以便在最小二乘意义下来说明数据中的最大可能的方差．这一

重要的向量穿过了数据点的最集中的部分，第一个特征向量为数据定义了一个最好的 1 因子模型．一般地讲，第一个因子说明了数据方差的主要部份，这个因子代表了一类可以说是在所有的指定上求平均值的常见的因子．

这样的计算手续一步一步地继续下去直至全部的 c 个轴都被确定位置．为找到每下一个轴的位置，必须利用两个条件：①尽可能多的方差被一个因子加以说明；②最新的轴与已被定位了的轴组相互正交．最紧接着的特征向量说明了已被第一个特征向量定义了的平均行为的方差．每一个连续的因子担负着数据中的总的方差的一个较小的部份．第二个主轴与第一个主轴正交．这一因子指向某一个方向，该方向说明尽可能多的没有被第一个因子说明过的方差．最初的这两个因子定义了一个平面，该平面概括了数据点的最大部份．该平面上的点被用两个坐标轴的一个 2 因子模型来加以说明．

与连续的较小的特征值相对应的特征向量越来越多地只说明一些“独特”的行为，这些行为相对地讲只与少数几个指定或甚至只是某单一的指定有关．在化学问题中，较小的特征值说明独特的行为，最终只是说明实验误差而已．第 c 个特征向量被定位于因子空间中以便说明最后那点点实验误差．c 个特征向量组成的完整集精确地说明了数据的每一部份，包括实验误差．

为了掌握因子的线索，我们将用括号来指明正被考虑的因子数目，例如 $d_{ik}(m)$ 是指用最前面 m 个主因子来复原所得的第 i 行第 k 列的数据点．因此写成

$$d_{ik}(m) = \sum_{j=1}^{m} r_{ij}c_{jk}, \tag{2.34}$$

式中，加和包含前面 m 个主因子．为得到第一个主因子，做以下处理．首先，我们定义残余误差 $e_{ik}(1)$ 为实验数据点 d_{ik} 和复原所得数据点 $d_{ik}(1)$ 之间的差异．用一个因子 (即第一个主因子) 时

$$e_{ik}(1) = d_{ik} - d_{ik}(1), \tag{2.35}$$

从式 (2.34) 可知

$$e_{ik}(1) = d_{ik} - r_{i1}c_{1k}. \tag{2.36}$$

为使残余误差为最小，应用最小二乘法. 根据我们所希望强调的指定的类型，将每一个残余误差的平方对行余因子或列余因子求导数. 如果是对行余因子求导，则在 c 列中进行加和；如果是对列余因子求导，则在 r 行中进行加和. 以强调行指定为例，应用下面的手续得到

$$\sum_{i=1}^{r} \frac{\mathrm{d}e_{ik}^2(1)}{\mathrm{d}c_{1k}} = 2c_{1k}\sum_{i=1}^{r} r_{i1}^2 - 2\sum_{i=1}^{r} r_{i1}d_{ik}, \tag{2.37}$$

按最小二乘原理，设这一加和等于 0，得到

$$\sum_{i=1}^{r} r_{i1}d_{ik} = c_{1k}\sum_{i=1}^{r} r_{i1}^2. \tag{2.38}$$

由于 k 从 1 到 c 变化，共有 c 个这种形式的方程. 用矩阵符号可表示如下

$$R_1'[D] = C_1'R_1'R_1, \tag{2.39}$$

与式 (2.28) 一致，我们用下式来定义 λ_1

$$\lambda_1 = R_1'R_1 = \sum_{i=1}^{r} r_{i1}^2, \tag{2.40}$$

将此式代入式 (2.39) 并取转置，我们发现

$$[D]^T R_1 = \lambda_1 C_1. \tag{2.41}$$

根据矩阵的运算性质，完整的数据阵可写成

$$[D] = R_1C_1' + R_2C_2' + \cdots + R_cC_c', \tag{2.42}$$

由于加和已包含 c 个特征向量，这一方程只影响到包括实验误差在内的完整的数据矩阵 $[D]$ 而不影响矩阵 $[D^{\ddagger}]$，后者涉及到仅用 n 个

主要特征向量的加和

$$[D^{\ddagger}]=[R^{\ddagger}][C^{\ddagger}]=[R_1\ R_2\ \cdots\ R_n]\begin{bmatrix}C_1'\\C_2'\\\cdots\\C_n'\end{bmatrix}. \tag{2.43}$$

通过用 C_1 来右乘式 (2.42)，设 $C_i'C_j=\delta_{ij}$，因特征向量是标准正交的，于是可得到

$$[D]C_1=R_1, \tag{2.44}$$

将式 (2.44) 代入式 (2.41)，得到

$$[D]^T[D]C_1=\lambda_1C_1, \tag{2.45}$$

依据前面所讲过的协方差阵的定义 (见式 (2.1))，现在可写出

$$[Z]C_1=\lambda_1C_1, \tag{2.46}$$

这一表达式可用以计算第一个主特征向量及其相应的特征值 λ_1. 这一点，在以后将会详细地加以讨论.

为了得到第二个主因子，我们考虑第二个残余误差

$$e_{ik}(2)=d_{ik}-d_{ik}(2), \tag{2.47}$$

从式 (2.34) 的定义，它可被用下式来表示

$$e_{ik}(2)=e_{ik}(1)-r_{i2}c_{2k}. \tag{2.48}$$

为使第二个主因子中的误差为最小，在保持 $e_{ik}(1)$ 不变的同时，我们对 $e_{ik}(2)$ 使用最小二乘法. 这样，便可得到一个与式 (2.37) 相似的表达式

$$\sum_{i=1}^{r}\frac{\mathrm{d}e_{ik}^2(2)}{\mathrm{d}c_{2k}}=2c_{2k}\sum_{i=1}^{r}r_{i2}^2-2\sum_{i=1}^{r}r_{i2}e_{ik}(1), \tag{2.49}$$

为使误差为最小，让这一加和等于 0，得出

$$\sum_{i=1}^{r}r_{i2}e_{ik}(1)=c_{2k}\sum_{i=1}^{r}r_{i2}^2, \tag{2.50}$$

因为有 c 个这样的方程式，以矩阵符号表示可写成

$$R_2'[E]_1 = C_2'R_2'R_2, \tag{2.51}$$

式中，$[E]_1$ 是一个由第一个残余误差组成的 $r \times c$ 矩阵，现在 λ_2 被定义为

$$\lambda_2 = R_2'R_2 = \sum_{i=1}^{r} r_{i2}^2, \tag{2.52}$$

从式中 (2.51) 和 (2.52)，可得到

$$[E]_1R_2 = \lambda_2C_2, \tag{2.53}$$

矩阵 $[E]_1$，可被写成

$$[E]_1 = [D] - R_1C_1' = R_2C_2' + R_3C_3' + \cdots + R_cC_c', \tag{2.54}$$

用 C_2 右乘式 (2.54)，记住特征向量是标准正交的，于是得到

$$[E]_1C_2 = R_2. \tag{2.55}$$

将此式代入式 (2.53)，给出

$$[E]_1^T[E]_1C_2 = \lambda_2C_2, \tag{2.56}$$

从式 (2.39), (2.40) 和 (2.54)，我们可写出

$$[E]_1^T[E]_1 = [D]^T[D] - \lambda_1C_1C_1', \tag{2.57}$$

第一个残余矩阵定义为

$$[\mathcal{R}]_1 = [Z] - \lambda_1C_1C_1', \tag{2.58}$$

从式 (2.56), (2.57) 和 (2.58)，可得出结论

$$[\mathcal{R}]_1C_2 = \lambda_2C_2, \tag{2.59}$$

同前所述，这个表达式可用来计算第二个主特征向量及其相应的特征值 λ_2.

解的数据矩阵为 $[D]$

$$[D]=\begin{bmatrix} 4 & 40 & -20 & 32 \\ 1 & -5 & 17 & -5 \\ 4 & 10 & 4 & 6 \\ 6 & 15 & 6 & 9 \\ 11 & 50 & -7 & 36 \\ 6 & -15 & 30 & -17 \\ 17 & 80 & -13 & 58 \\ 22 & 115 & -26 & 85 \\ 14 & 20 & 26 & 8 \\ 13 & -5 & 43 & -13 \end{bmatrix}. \tag{2.64}$$

先按式 (2.19) 构造某协方差矩阵 $[Z]$

$$[Z]=[D]^T[D]=\begin{bmatrix} 1364 & 4850 & 212 & 3294 \\ 4850 & 24725 & -5230 & 18195 \\ 212 & -5230 & 4820 & -4674 \\ 3294 & 18195 & -4674 & 13573 \end{bmatrix}. \tag{2.65}$$

为进行迭代，参照式 (2.46)

$$[Z]C_1=\lambda_1 C_1,$$

这里，C_1 是第一个特征向量，λ_1 是其对应的特征值. 作为第一级近似，我们随意设

$$C_1'=(0.10000\ 0.10000\ 0.10000\ 0.10000)$$

是一个归一化过的向量. 按照式 (2.46)，用协方差阵 $[Z]$ 左乘这个向量

$$\begin{bmatrix} 1364 & 4850 & 212 & 3294 \\ 4850 & 24725 & -5230 & 18195 \\ 212 & -5230 & 4820 & -4674 \\ 3294 & 18195 & -4674 & 13573 \end{bmatrix}\begin{bmatrix} 0.10000 \\ 0.10000 \\ 0.10000 \\ 0.10000 \end{bmatrix}=\begin{bmatrix} 972 \\ 4254 \\ -487.2 \\ 3038.8 \end{bmatrix},$$

对上式右边所得的列向量进行归一化. 用列向量中全部元素的平方

和的平方根去除每一个元素便可达到归一化的目的

$$\begin{bmatrix} 972 \\ 4254 \\ -487.2 \\ 3038.8 \end{bmatrix} = 5339.754 \begin{bmatrix} 0.18203 \\ 0.79667 \\ -0.09124 \\ 0.56909 \end{bmatrix}.$$

上式中，5339.754 是一个归一化常数，可看作是 λ_1 的一个近似值. 作为第二级近似，我们采用归一化过的向量，即设

$$C_1' = (0.18203 \quad 0.79667 \quad -0.09124 \quad 0.56909),$$

然后，再用 $[Z]$ 去左乘这一新的向量

$$\begin{bmatrix} 1364 & 4850 & 212 & 3294 \\ 4850 & 24725 & -5230 & 18195 \\ 212 & -5230 & 4820 & -4674 \\ 3294 & 18195 & -4674 & 13573 \end{bmatrix} \begin{bmatrix} 0.18203 \\ 0.79667 \\ -0.09124 \\ 0.56909 \end{bmatrix} = \begin{bmatrix} 5967.36 \\ 31412.19 \\ -7227.68 \\ 23245.66 \end{bmatrix},$$

对上式右边所得的列向量进行归一化，得

$$\begin{bmatrix} 5967.36 \\ 31412.19 \\ -7227.68 \\ 23245.66 \end{bmatrix} = 40186.26 \begin{bmatrix} 0.14849 \\ 0.78167 \\ -0.17985 \\ 0.57845 \end{bmatrix}.$$

上式右边得到的列向量是 C_1 的一个更好近似，归一化常数 40186.26 是 λ_1 的一个更好的近似. 上述过程一直重复下去，每一次都会得到 C_1 和 λ_1 的更好的近似，直至 C_1 的各元素收敛为常数并满足式 (2.46)，最后 (实际上总共只需要做 8 次) 得到

$$\lambda_1 = 40416.13,$$

和

$$C_1' = (0.144633 \quad 0.779523 \quad -0.189719 \quad 0.579165).$$

为了得到第二个特征向量，我们按式 (2.58) 计算第一个残余矩阵

$$[\mathcal{R}]_1 = [Z] - \lambda_1 C_1 C_1'.$$

我们的计算如下

$$\lambda_1 C_1 C_1' = 40416.13 \begin{bmatrix} 0.144633 \\ 0.779523 \\ -0.189719 \\ 0.579165 \end{bmatrix} \begin{bmatrix} 0.144633 \\ 0.779523 \\ -0.189719 \\ 0.579165 \end{bmatrix}'$$

$$= \begin{bmatrix} 845.45 & 4656.71 & -1109.01 & 3385.52 \\ 4556.71 & 24559.12 & -5977.15 & 18246.176 \\ -1109.01 & -5977.15 & 1454.7 & -4440.86 \\ 3385.52 & 18246.76 & -4440.86 & 13556.85 \end{bmatrix},$$

将矩阵 $[Z]$ 减去此矩阵便可生成第一个残余矩阵

$$[\mathcal{R}]_1 = \begin{bmatrix} 518.55 & 293.29 & 1321.01 & -91.52 \\ 293.29 & 165.88 & 747.15 & -51.76 \\ 1321.01 & 747.15 & 3365.29 & -233.14 \\ -91.52 & -51.76 & -233.14 & 16.15 \end{bmatrix}.$$

根据式 (2.59)，用同计算第一个特征向量相似的迭代操作，即可求得 λ_2 和 C_2. 作为开始值，可对 C_2 的各元素随意取值，如

$$C_2' = (0.10000 \quad 0.10000 \quad 0.10000 \quad 0.10000),$$

只要进行 3 次迭代，最后便可得到

$$\lambda_2 = 4065.868,$$

和

$$C_2' = (0.357122 \quad 0.201986 \quad 0.909776 \quad -0.630266).$$

按式 (2.61)

$$[\mathcal{R}]_2 = [\mathcal{R}]_1 - \lambda_2 C_2 C_2'.$$

计算得到

$$[\mathcal{R}]_2 = \begin{bmatrix} 0.61 \times 10^{-4} & 2.1 \times 10^{-4} & 1.2 \times 10^{-4} & -1.4 \times 10^{-4} \\ 2.1 \times 10^{-4} & -0.61 \times 10^{-4} & -1.2 \times 10^{-4} & -11.94 \times 10^{-4} \\ 0 & -1.2 \times 10^{-4} & 4.98 \times 10^{-4} & 0.92 \times 10^{-4} \\ 1.1 \times 10^{-4} & -11.98 \times 10^{-4} & 0.92 \times 10^{-4} & -16.4 \times 10^{-4} \end{bmatrix}$$

这一矩阵实质上等于 0 ，那些小的有限值是由于计算过程的取舍造成的. 对于这一实例数据矩阵来说，分解过程到此已全部结束了. 对于其他数据矩阵，如果这时 $[\mathcal{R}]_2$ 实质上尚未等于零，则分解过程应继续进行下去，以取得 C_3 ， C_4 ， $\cdots$ 和 C_c 以及它们对应的特征值 λ_3 ， λ_4 ， $\cdots$ 和 λ_c.

以上介绍的方法亦称为乘幂法. 由于只有少数几个最大的特征值及其对应的特征向量被加以计算，故计算机时和储存空间都比较节省.

2.3.4 JACOBI 法

在化学问题中，由实验数据矩阵构造成的 c 阶协方差矩阵 $[Z]$ 一般都为实对称矩阵，满足厄米条件. 对于给定的 $[Z]$，存在 c 个叫做特征值的实常数： $\lambda_1, \lambda_2, \cdots, \lambda_c$ ，而且，对应于全部特征值也存在着 c 个叫做特征向量的实向量： C_1 ， C_2 ， $\cdots$ ， C_c.

对于 $[Z]$ 的分解，虽然有各种方法 (如前面已介绍过的主因子分析，即乘幂法)，但由于在实践中，矩阵 $[Z]$ 的阶数都较高，故为了避免高阶特征方程求根的麻烦，求得精度较高的结果且保证所求得的特征向量有很好的正交性， JACOBI 法这一经典的技术在实际计算中还是很受欢迎的.

矩阵的特征值在相似变换下保持不变，即相似矩阵有共同的特征值. 这是 JACOBI 法的矩阵数学前提.

设矩阵 $[A]$ 为 $c \times c$ 矩阵，如存在非奇异矩阵 $[M]$ ，使

$$[A] = [M]^{-1}[Z][M], \tag{2.66}$$

则称矩阵 $[A]$ 和 $[Z]$ 相似，并称 $[M]$ 是化 $[Z]$ 为 $[A]$ 的相似变换. 即 $[A]$ 和 $[Z]$ 具有共同的特征值. 如 $[M]$ 为 $c \times c$ 矩阵，且它的转置矩阵等于它的逆矩阵，即

$$[M]^T = [M]^{-1}, \tag{2.67}$$

则称 $[M]$ 为正交矩阵，有

$$[A]=[M]^T[Z][M], \tag{2.68}$$

这一操作称为矩阵 $[Z]$ 的正交变换，这同时也是一种相似变换．任意实对称矩阵 $[Z]$ 总可以通过正交相似变换化为对角阵．一个对角阵的特征值即为对角元． JACOBI 法是一种旋转法，是一种用平面旋转矩阵所构成的正交相似变换将对称矩阵化为对角矩阵的经典方法，实际上， JACOBI 法就是对式 (2.68) 进行一系列的操作．先选一个 $[M]$，使 $[Z]$ 的一个非对角元素变成零，此时 $[Z]$ 变成 $[A]$；然后，再选第二个 $[M]$，使 $[Z]$ 的另一个非对角元素变成零，继续这种操作直至产生的被变换的矩阵成为对角阵．例如，要使矩阵

$$[Z]=\begin{bmatrix} 1 & 2 & -1 & 3 \\ 2 & -3 & 4 & 6 \\ -1 & 4 & 3 & 0 \\ 3 & 6 & 0 & 13 \end{bmatrix}$$

中最大的非对角元素 ($z_{24}=z_{42}=6$) 变为零，选择正交矩阵

$$[M_1]=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta & 0 & \sin\theta \\ 0 & 0 & 1 & 0 \\ 0 & -\sin\theta & 0 & \cos\theta \end{bmatrix}.$$

假设 i 和 j 是要使之为零的非对角元素的行下标和列下标，$[Z_1]$ 是第一次正交变换的结果，即

$$[Z_1]=[M_1]^T[Z][M_1], \tag{2.69}$$

那么，$[Z_1]$ 的元素可按下列公式计算

$$\left.\begin{aligned}
z_{kl}^{(1)} &= z_{kl}, \quad (k,l \neq i,j),\\
z_{il}^{(1)} &= z_{il}\cos\theta - z_{jl}\sin\theta,\\
z_{jl}^{(1)} &= z_{il}\sin\theta + z_{jl}\cos\theta,\\
z_{ki}^{(1)} &= z_{ki}\cos\theta - z_{kj}\sin\theta,\\
z_{kj}^{(1)} &= z_{ki}\sin\theta + z_{kj}\cos\theta,\\
z_{ii}^{(1)} &= z_{ii}\cos^2\theta + z_{jj}\sin^2\theta - 2z_{ij}\sin\theta\cos\theta,\\
z_{jj}^{(1)} &= z_{ii}\sin^2\theta + z_{jj}\cos^2\theta + 2z_{ij}\sin\theta\cos\theta,\\
z_{ij}^{(1)} &= (z_{ii} - z_{jj})\sin\theta\cos\theta + z_{ij}(\cos^2\theta - \sin^2\theta).
\end{aligned}\right\} \tag{2.70}$$

因为这次变换的目的是要使 $z_{ij}^{(1)}$ 为零，即

$$z_{ij}^{(1)} = (z_{ii} - z_{jj})\sin\theta\cos\theta + z_{ij}(\cos^2\theta - \sin^2\theta) = 0, \tag{2.71}$$

因为 $\sin\theta\cos\theta = 1/2\sin 2\theta$, $\cos^2\theta - \sin^2\theta = \cos 2\theta$, 所以，式 (2.71) 可改写成

$$1/2(z_{ii} - z_{jj})\sin 2\theta + z_{ij}\cos 2\theta = 0, \tag{2.72}$$

即

$$\mathrm{tg}2\theta = \frac{-2z_{ij}}{z_{ii} - z_{jj}}. \tag{2.73}$$

确定 θ 之后，可求得 $\sin\theta$ 和 $\cos\theta$ 的值，第一次变换便可以顺利的完成.

用 $[Z_1]$ 取代 $[Z]$，分别选用正交阵 $[M_2]$，$[M_3]$，$\cdots$，重复上述过程求出矩阵 $[Z_2]$，$[Z_3]$，$\cdots$. 假定我们在最后求出矩阵 $[Z_x]$，且 $\max\limits_{p \neq q} |Z_{pq}^{(x)}| \leq \epsilon$，即意味着 JACOBI 法已收敛，并已得到近似的对角矩阵. 此时，$[Z_x]$ 的对角元就是所求的特征值. 矩阵的乘积 $[S] = [M_1][M_2][M_3]\cdots[M_x]$，也是正交阵. 且矩阵 $[S]$ 的第 i 列即为 $[Z]$ 的第 i 个特征向量.

然而，利用式 (2.73) 时存在一些问题：使用三角函数引入截断误差；还要提防 $A_{pp} = A_{qq}$ 这样一种特殊情况. 通过下列三角函数代换可克服这些问题.

设

$$\lambda = -z_{ij}, \quad \mu = \frac{z_{ii} - z_{jj}}{2},$$

$$d = \text{sign}(\mu)\frac{\lambda}{\sqrt{\lambda^2 + \mu^2}},$$

其中，sign 是一个符号函数，当 $\mu \geq 0$ 时，取值 $+1$; $\mu < 0$ 时，取值 -1. 则有

$$\sin\theta = \frac{d}{\sqrt{2(1 + \sqrt{1 - d^2})}}, \tag{2.74}$$

$$\cos\theta = \sqrt{1 - \sin^2\theta}. \tag{2.75}$$

用式 (2.74) 和 (2.75) 替代式 (2.70) 中的 $\sin\theta$ 和 $\cos\theta$，则可顺利进行变换.

现在，让我们回到前面提出的例子上去，当 $z_{24} = z_{42} = 6$ 时，$i = 2$，$j = 4$，则有

$$\lambda = -z_{24} = -6, \quad \mu = \frac{z_{22} - z_{44}}{2} = \frac{-3 - 13}{2} = -8,$$

代入 $\text{Sign}(\mu) = -1$，得

$$d = (-1)\frac{-6}{\sqrt{36 + 64}} = (-1)\frac{-6}{10} = 0.6,$$

$$\sin\theta = \frac{d}{\sqrt{2(1 + \sqrt{1 - (0.6)^2})}} = \frac{0.6}{\sqrt{3.6}} = 0.3162,$$

$$\cos\theta = \sqrt{1 - (0.3162)^2} = 0.9487.$$

将 $\sin\theta$ 和 $\cos\theta$ 的值代入式 (2.70)，就可算得第一次变换的结果为

$$[Z_1] = \begin{bmatrix} 1 & 0.95 & -1 & 3.48 \\ 0.95 & -5 & 3.80 & 0 \\ -1 & 3.80 & 3 & 1.26 \\ 3.48 & 0 & 1.26 & 15 \end{bmatrix}.$$

现在 $[Z_1]$ 中最大的非对角元素是 $z_{23}^{(1)}=z_{32}^{(1)}=3.80$，用 $Z^{(1)}$ 替代式 (2.70) 中的 Z，算出相应的 $Z^{(2)}$，进行第二次变换，使这一元素为零，得到

$$[Z_2]=\begin{bmatrix} 1 & 1.25 & -0.58 & 3.48 \\ 1.25 & -6.51 & 0 & -0.47 \\ -0.58 & 0 & 4.51 & 1.17 \\ 3.48 & -0.47 & 1.17 & 15 \end{bmatrix}.$$

从以上变换可以看到，只有最大的非对角元素所在的行和列元素才受其相应的变换的影响，其他元素则保持不变. 而 $z_{24}^{(1)}=z_{42}^{(1)}$，原来为 0，第二次变换后变成 $z_{24}^{(2)}=z_{42}^{(2)}=-0.47$. 但不需担心，JACOBI 法具有较好的收敛性. 进行 16 次变换后，便得到下面近似对角阵

$$[Z_{16}]=\begin{bmatrix} 15.91 & -1.1\times10^{-9} & 0 & -1.1\times10^{-14} \\ -1.1\times10^{-9} & 4.584 & -1.7\times10^{-13} & 3.8\times10^{-22} \\ 0 & -1.7\times10^{-13} & 0.2775 & -1.0\times10^{-23} \\ -1.1\times10^{-14} & 3.8\times10^{-22} & -1.0\times10^{-23} & -6.770 \end{bmatrix},$$

此时，$[Z_{16}]$ 的对角元便是所求的特征值

$$\lambda_1=15.91,\quad \lambda_2=4.584,\quad \lambda_3=0.2775,\quad \lambda_4=-6.770.$$

从各次正交变换矩阵的乘积，便可求得对应于特征根的特征向量

$$[M_1][M_2][M_3]\cdots[M_x]=[S],$$

$[S]$ 中的各列向量正好是属于 $[Z]$ 的各个特征值的特征向量. 有

$$[S]^{-1}[Z][S]=[Z_x]\approx[D], \tag{2.76}$$

$$[Z][S]=[S][D], \tag{2.77}$$

$[D]$ 为对角矩阵. 如果把式 (2.76) 中 [S] 的第 l 列向量记为 C_l，$[D]$ 对角线上第 l 个元素记为 λ_l，则 $[Z]C_l=\lambda_l C_l$. 可见 $[S]$ 中的每一列元素就为对应 $[D]$ 上特征值的特征向量.

清楚地加以讨论过. 式 (2.27) 可用矩阵形式表示成

$$[Z] = [C]^T[\lambda][C], \tag{2.78}$$

式中

$$[C] = \begin{bmatrix} C_1' \\ C_2' \\ \cdots \\ C_c' \end{bmatrix} = [C_1'\ C_2'\ \cdots\ C_c]^T. \tag{2.79}$$

由此可见，从迭代过程结果所得到的特征向量构成了列矩阵 $[C]$ 的相应的行. 注意，我们在这里考虑由 c 个特征向量组成的完整集，它说明了全部的数据，包含实验误差在内. 因为 $[C]$ 是标准正交的，它的转置等于它的逆

$$[C]^T = [C]^{-1}, \tag{2.80}$$

重排式 (2.78) 得

$$[C][Z][C]^{-1} = [\lambda]. \tag{2.81}$$

式 (2.81) 右边的矩阵 $[\lambda]$ 是一个含有特征值作为对角元的对角矩阵. 在这种情形下，矩阵 $[C]$ 可被当作是一个对角化矩阵. 可见，对角化矩阵等于列矩阵. 由于它是由特征向量组成的，它普遍地被称为“向量矩阵”.

2.3.6 计算行矩阵

重排式 (2.9) 并联系到式 (2.80) 可得到

$$[R] = [D][C]^{-1} = [D][C]^T, \tag{2.82}$$

计算得到 $[C]$ 后，按上式所示的乘法，可以算得行矩阵 $[R]$. 它的每一个元素代表一个行指定点在相应的特征向量上的“投影”.

在更进一步处理之前，让我们详细地来讨论行矩阵，它的列是相互正交的. 关于这一点的证明是这样的：我们按先后顺序来分别应用式 (2.82), (2.1), (2.80) 和 (2.81)，得

$$\begin{aligned} [R]^T[R] &= ([D][C]^T)^T([D][C]^T) = [C][D]^T[D][C]^T \\ &= [C][Z][C]^{-1} = [\lambda]. \end{aligned} \tag{2.83}$$

从这一结果，我们得出结论

$$R_j' R_j = \lambda_j, \tag{2.84}$$

和

$$R_i' R_j = 0, \tag{2.85}$$

式中 R_j 是行矩阵 $[R]$ 的一个列向量，它与特征值 λ_j 有联系．结合式 (2.28) 和 (2.84) 得出结果

$$R_j' R_j = \lambda_j = \sum_{i=1}^{r} r_{ij}^2. \tag{2.86}$$

正如所期待的那样，这一结果与式 (2.40) 和 (2.52) 相一致．

从式 (2.86) 可见到，特征值是行指定在一给定的特征向量上投影的平方加和．由于在 R_j 中存在着象数据矩阵的行数一样多的元素，从式 (2.86) 可以明白 R_j 中的每一个元素代表该行指定在特征向量轴 C_j 上的投影或得分．鉴于此，行矩阵一般地被称为投影矩阵或得分矩阵．与 C_j 不一样，R_j 是不归一化的，但可通过用特征值 λ_j 的平方根去除 R_j 中的每一个元素而达到归一化的目的．

2.3.7 因子的数目和复原

在完成分解步骤之后，我们的任务就是要去发现在全部所得的 c 个因子中到底有多少个因子是具有物理或化学意义的．全部所得的抽象因子可以分成两类：①包含有 n 个因子，它们表明数据真正的可量测的性质；②包含有 $c-n$ 个因子，它们完全与实验误差有关．通过从最初的抽象解中剔除第二类因子，我们可“压缩”因子模型以便仅体现那些具有物理意义的因子．

压缩后，式 (2.9) 变成

$$[D] = [R]_{n,\mathrm{PFA}}[C]_{n,\mathrm{PFA}}. \tag{2.87}$$

式 (2.87) 准确地表达了经过适当计算的抽象解．这是以后所进行的有关因子分析计算的基础．

化学工作者从因子分析的应用中所获取的第一种收益是正确的因子数目的确定．真实的因子数目则是数据的真实复杂性的量测，而这一类信息通过其他方法是很少能得到的．

如果实验误差已知，则许多方法，如马上将要叙述的数据复原手续便可用来找到因子的数目．如果数据中的误差未知，可用一些特殊的数字方法来估算因子的数目．确定真实因子数目的经验的和理论的方法将于第七章中详细讨论．

逐步地进行的抽象复原法可被用来推断因子的正确数目．复原的每一步包括以下的计算和比较

$$[R]_{j,\mathrm{PFA}}[C]_{j,\mathrm{PFA}} = [D]_j \stackrel{?}{=} [D], \tag{2.88}$$

式中，$[R]_{j,\mathrm{PFA}}$ 和 $[C]_{j,\mathrm{PFA}}$ 是以 j 个最重要的特征向量为依据的抽象矩阵，$[D]_j$ 是应用最前面 j 个抽象因子进行复原而得到的数据矩阵，$[D]$ 是原始数据矩阵．在做第一步复原时，只有一个，即最重要的那个因子 ($j=1$) 被采用；在复原的第二步，第一个和第二个最重要的因子 ($j=2$) 被同时采用，照此类推，直至在最后的复原步骤中所有 c 个因子一起被采用．

当正确的因子数目被采用时 ($j=n$)，那么，经复原得到的矩阵 $[D]_n$ 应在实验误差范围内与原始数据矩阵相等．作为 c 个因子中的一个完整集，一个肯定的数目 n，被要求用来在实验误差范围内复原数据．如果太少的因子在抽象因子分析模型中被采用，则不能在具有足够的精确度的范围内复原数据；如果大多的因子被采用，则多余的因子将会产生实验误差，因而也就毫无用处．采用 n 个因子 (即 n 个特征向量) 来决定因子空间，故有

$$[D] \cong [D^{\ddagger}] = [R^{\ddagger}][C^{\ddagger}] = [R_1\ R_2\ \cdots\ R_n]\begin{bmatrix} C_1' \\ C_2' \\ \cdots \\ C_n' \end{bmatrix}. \tag{2.89}$$

这里，$[R^{\ddagger}]$ 和 $[C^{\ddagger}]$ 分别称为抽象行矩阵和抽象列矩阵．存在 n 个

与行矩阵对应的列和 n 个与列矩阵对应的行，因子空间是 n 维的．这一过程称为短路复原．

在化学问题所要求得到的信息只是因子数目时，预备、分解和复原步骤便构成了完整的因子分析流程．

通过式 (2.89)，我们的确可以在实验误差范围内复原数据矩阵，但我们不能对结果所得的矩阵赋予任何物理意义，因为它们仅代表数学的解．由于在这一过程中只采用空间的数学抽象因子，故我们将这整个过程称为抽象因子分析．

在化学问题中，所有的量测都不可避免地含有实验误差．以随机方式进入数据的误差产生了多余的特征向量，而这样的特征向量是没有真实意义的．在因子分析方法中对它们的不必要保留既增加了空间的维数，同时也将产生其精度远不是所期望的预报．

当然，由于实验误差的存在，辨认因子空间的维数并不是一件轻而易举的任务．为此而长期进行了有关判据的研究工作．关于此，将于第七章中详细讨论．其中，有一种简单判据是以比较原来的数据矩阵与用短路复原步骤预测得到的矩阵做为基础的，为方便起见，安排在这里加以讨论．

为了确定因子空间的维数，我们从与最大的特征值 λ_1 相对应的特征向量 C_1 开始．这一特征向量是最重要的，它说明了数据中的最大的偏差，关于这一点，从式 (2.28) 可很好地加以说明．该式表明特征值等于行指定在特征向量轴上的得分的平方加和，即 $\lambda_j = \sum r_{ij}^2$，式中，加和涉及全部的元素．现在，执行下面的矩阵乘法

$$[D]_1 = [R_1][C_1'], \tag{2.90}$$

式中，R_1 和 C_1' 是与 λ_1 相对应的各自的向量．将这时计算所得的矩阵 $[D]_1$ 同原来数据矩阵 $[D]$ 进行比较，如果它们之间的一致性不在实验误差范围内的话，我们用下面一个最重要的特征向量来继续分析

$$[D]_2 = [R_1 \ \ R_2]\begin{bmatrix} C_1' \\ C_2' \end{bmatrix}, \tag{2.91}$$

如尚达不到一致性，我们依次采用下一个最重要的特征向量并继续分析下去，直至能满意地复原数据．这样，我们会发现

$$[D]_n = [R_1 \quad R_2 \cdots R_n] \begin{bmatrix} C_1' \\ C_2' \\ \cdots \\ C_n' \end{bmatrix} = [D^{\dagger}] \cong [D], \tag{2.92}$$

式中，C_n 是最后一个需用来复原数据的特征向量轴．因子空间是 n 维的．

需要留意的是，如果 n 大于 r 或 c (数据矩阵的行数或列数)，那么，要么就是我们没有引入足够的数据 (即所用数据不旋转因子空间)，要么就是所有的数据不能用相同的因子来表达 (即存在独特的大自由度)．另一方面，如果 n 小于 r 或 c，则因子分析的这一步骤便告完成．

在实践中，最好应力求使 r 和 c 中的较小者至少应为 n 的两倍．增加数据阵的行或列数目可达到这一目的．提醒这一点，目的在于确保比未知的余因子有多得多的数据点．

到此，我们已采用压缩因子空间来完成图 2.2 中的短路复原循环，因子分析在抽象意义上已全部完成了．

2.4 向量的诠释

从向量的观点来考虑问题，可以帮助我们对因子分析的总的操作细节有一个深刻的理解．将数据矩阵的列当作向量，按式 (2.1) 求每一对的列的点积即可得到协方差矩阵的元素．如果是在相关阵的情况下，则在取点积之前应该先对数据阵的每一列进行归一化．相关阵的每一个元素代表两个有关的数据列向量之间的夹角的余弦，相关阵的对角元素均为 1，因为这是通过取各向量自身的点积而获得的．

如果 n 个特征向量被要求用来复原数据矩阵，则全部列向量分布在 n 维空间中，要求要有 n 个正交参比轴．为了更好地理解这一

点，我们来考虑这样一个例子 —— 一个归一化了的数据矩阵，它由产生于两个因子的五个数据列 C_1，C_2，C_3，C_4 和 C_5 组成．得到该矩阵的相关阵为 $[Z]_N$

$$[Z]_N =$$

$$\begin{bmatrix} 1.00000 & 0.00000 & 0.80902 & 0.40674 & -0.70711 \\ 0.00000 & 1.00000 & 0.58779 & 0.91355 & 0.70711 \\ 0.80902 & 0.58779 & 1.00000 & 0.86603 & -0.15634 \\ 0.40674 & 0.91935 & 0.86603 & 1.00000 & 0.35837 \\ -0.70711 & 0.70711 & -0.15634 & 0.35837 & 1.00000 \end{bmatrix}. \tag{2.93}$$

此相关矩阵中的元素是数据列向量之间夹角的余弦．据此，我们可以给出原数据阵中各列向量之间的向量关系，如图 2.4 所示我们发现全部 5 个向量分布在一个共同的平面上，问题是二维的，即仅涉及两个因子．

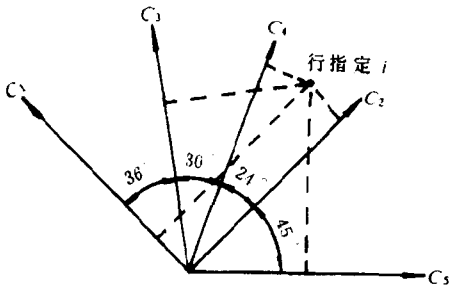


图 2.4　生成式 (2.93) 的相关阵所用的数据列向量的向量关系．通过一个行指定的点在各对应列指定向量 C_1，C_2，C_3，C_4 和 C_5 上的垂直投影即可得到数据点

图 2.4 中的每个向量轴对应于已归一化了的数据阵的一个列指定，用平面上的一个点表示已归一化了的数据阵的一个行指定 (见图中的行指定 i 点)，可以这样来获得与一个给定的行和列有关的数据的值：通过该行指定点划一条与相应的列向量垂直的线与该列向量相交，然后，沿着向量读出原点到交叉点的距离，这些投影便是归一化了的数据值，因为向量轴代表归一化了的数据列．从广义上说，

投影至任何轴上所得的点的值就被叫做投影或因子成分，在传统因子分析中，这些投影被叫做“得分”.

如果图 2.4 中，有一个向量不分布在那个平面里，则空间便是三维的. 这时，要求有 3 个因子来说明数据，在因子空间中就要求用 3 个轴来对数据点进行定位. 相关矩阵的秩就是 3. 参比轴的选择不完全是随意的. 例如，如果向量 1 不在平面上，则可用向量 1 和其他 4 个向量中的任两个来作为参比轴. 不能用向量 2, 3, 4 和 5 来作为参比轴，因为它们都分布在同一平面中，不能旋转三维空间.

在许多问题中，因子空间超过三维，我们不能用二维图来描述这类多维情况. 但是，借助于计算机来提取出必要的信息是可能的. 应用因子分析技术可以确定因子空间的准确维数. 因子分析所获得的特征向量旋转空间但与数据向量并不相一致，它们只定义所有实验数据共存的因子空间.

当式 (2.93) 中的相关矩阵被用于图 2.2 所示的分解步骤时，就会得到两个相互正交的特征向量 V_1 和 V_2 以及和它们相对应的特征值 λ_1 和 λ_2

$$\left.\begin{aligned} V_1 = \begin{bmatrix} v_{11} \\ v_{12} \\ v_{13} \\ v_{14} \\ v_{15} \end{bmatrix} = \begin{bmatrix} 0.30544 \\ 0.50260 \\ 0.54253 \\ 0.58338 \\ 0.13941 \end{bmatrix}, \quad V_2 = \begin{bmatrix} v_{21} \\ v_{22} \\ v_{23} \\ v_{24} \\ v_{25} \end{bmatrix} = \begin{bmatrix} -0.58845 \\ 0.35763 \\ -0.26583 \\ 0.08736 \\ 0.66898 \end{bmatrix}; \end{aligned}\right\} \tag{2.94}$$

$$\lambda_1 = 2.89104,$$

$$\lambda_2 = 2.10896.$$

定义特征向量的系数 v_{jk} 也衡量了每个特征向量在每一数据列上的重要性. 大体上讲，5 个数据列向量中的每一个可应用特征向量和特征值通过下式来表达

$$C_k = \sum_{j=1}^{n} \sqrt{\lambda_j} v_{jk} V_j. \tag{2.95}$$

着 ΔH_i 轴方向的距离来表示.

2.5 主因子分析解的变换

在大多数化学问题中，主要的目的还是要对因子性质有更深一步的了解. 所以，在复原步骤之后，因子分析过程不能就此结束，那时，我们所获得的主因子解只是纯粹的数学上的解，是基础的解，不是唯一的解. 抽象形式下的行和列因子并不具有任何明确的物理或化学意义，通过变换参数轴，可将主因子变换成在化学上可辨认的有实际意义的因子，这才是因子分析的最重要的收获.

变换就是通过寻找一个变换矩阵 $[T]$ 和它的逆矩阵 $[T]^{-1}$，用它们对主因子解进行如下的变换

$$\begin{aligned}[D] &= [R]_{\mathrm{PFA}}[C]_{\mathrm{PFA}} = \{[R]_{\mathrm{PFA}}[T]\}\{[T]^{-1}[C]_{\mathrm{PFA}}\} \\ &= [R]_{\mathrm{trans}}[C]_{\mathrm{trans}},\end{aligned} \tag{2.97}$$

通过合适的变换后，所得到的因子会变得与数据阵中的行或列指定的性质更加相匹配.

有两种明显不同的途径 (目标检验和抽象旋转) 被用来变换主因子分析解. 目标检验是一种唯一能每次检验一个潜在因子的方法，对于解决化学问题，它已显示出独特的优点. 它主要涉及两类因子：典型因子和基础因子 (总称为真实因子). 我们将在第三章中对它进行详细的讨论，本节着重对有关抽象旋转的内容作一般介绍. 所谓旋转，是赋予那些用来变换主因子分析抽象矩阵成为其他抽象矩阵的一类技术的总称，它主要涉及两类因子：主因子和旋转了的因子 (它们总称为抽象因子).

在某些限定内，我们可以变换轴并找到服从式 (2.9) 的解. 例如，假定有一些点，它们处于同一平面上，当然，它们的位置可用该平面上的两个互不相同的轴的坐标来指明. 这些轴可在平面上自由地旋转，原则上，存在着无穷多套可用来定义该平面并可确定数据点的轴. 与此相似，只要我们能维持这些轴的独特性，只要它们能充

分地旋转空间，则我们在因子分析中便可以旋转这些轴，特别是，我们希望变换这些轴使得它们能与行指定的基本结构参数相匹配.

执行下列数学操作即可完成轴的变换

$$[\overline{R}] = [R^{\ddagger}][T], \tag{2.98}$$

式中，$[T]$ 是一个 $n \times n$ 变换矩阵，$[\overline{R}]$ 是在新的坐标系中的行矩阵，$[R^{\ddagger}]$ 是抽象行矩阵.

变换矩阵的逆被用来在新的坐标系下确定列矩阵. 先将上式重排如下

$$[R^{\ddagger}] = [\overline{R}][T]^{-1}, \tag{2.99}$$

将此式代入式 (2.43)，得

$$[D^{\ddagger}] = ([\overline{R}][T]^{-1})[C^{\ddagger}] = [\overline{R}][T]^{-1}[C^{\ddagger}] = [\overline{R}][\overline{C}], \tag{2.100}$$

式中，$[\overline{C}]$ 为新坐标系下的列矩阵，被表达为

$$[\overline{C}] = [T]^{-1}[C^{\ddagger}]. \tag{2.101}$$

由此可见，通过对坐标轴的适当变换，发现一个可用化学或物理术语来诠释的行矩阵是可能的. 旋转一组轴可以通过无穷多的位置，所以，从因子分析结果可以得到无穷多数目的可能解，然而，只有轴的某些旋转才能产生出与可辨别的参数相对应的因子.

在因子分析的传统工作中，人们更普遍使用术语“旋转”而不是“变换”. 当对于因子的真实来源只有很少或是根本没有任何信息时，人们便采用抽象旋转，其最终目的是要提出具有简单因子结构的有意义的因子. 所有的旋转技术都努力确定这样一组轴，使得尽可能多的行指定点分布在这最终得到的因子轴附近，而仅存少数的点余留在被旋转的轴之间.

有许多获得旋转矩阵的方法，它们都以某些直观的因子空间判据，如简单结构等为依据，总的来说，可将它们分成两种类型：正交旋转和斜旋转. 正交旋转保持着从因子分析所得到的原来的向量集

之间的角的关系，如Quartimax 和方差最大等技术都属于此类. 斜旋转不保持向量间的角的关系，如Quartimin, Oblimax, Covarimin 和 Promax 等技术都属于此类.

简单地介绍一下上面列举的几种比较流行的抽象旋转方法对我们将来较详细地去讨论目标变换是会有一些帮助的.

Quartimax 是一种正交旋转技术，它保持着向量间的角的关系，其主要原理可通过一个二维例子来予以说明. 因为是正交轴被旋转，所以，一根轴逼近一个数据点，该点在此轴上的投影 (即载荷) 就增加，与此同时，该数据点离开另一根轴，它在这轴上的投影就减小. 这种技术主要是搜寻一组正交轴，这组轴能将点群中的数据点集中在每一根轴的附近，使每一个数据点在每一根轴上不是有高的载荷就是有低的载荷. 根据 Harman 的推导，通过对主因子向量轴的旋转，使得Quartimax 函数 (Q) 为最大即可达到上述目的

$$Q = \sum_{j=1}^{n}\sum_{k=1}^{c}\lambda_j^2\tilde{c}_{jk}^4, \tag{2.102}$$

式中，λ_j 是第 j 个特征值，$\tilde{c}_{jk}$ 是在旋转完成后第 k 个数据列向量在第 j 根轴上的载荷，加和涉及 c 个数据列和要求用来确定因子空间的 n 个因子. 这一技术的缺点是有“超载”第一个因子的倾向，其结果产生了一个大的普通的因子和许多小的次要的因子.

方差最大 (Varimax) 技术试图克服Quartimax 的缺点. 顾名思义，方差最大旋转就是使旋转后所得到的因子载荷矩阵的各列在保持彼此正交的前提下，其元素的平方后的总的方差 V 尽可能的大. 用数学公式表示，那就是使

$$V = \sum_{j=1}^{n}[1/c\sum_{k=1}^{c}(\lambda_j\tilde{c}_{jk}^2)^2 - 1/c^2(\sum_{k=1}^{c}\lambda_j\tilde{c}_{jk}^2)^2] \tag{2.103}$$

为最大. 式中，c 为样品数目，n 为因子的数目. 这样做的最终效果是使因子载荷矩阵的每一列元素按其平方值来讲，要么尽可能的

大，要么尽可能的小．问题实际上可归结为寻找一个 $n \times n$ 的正交矩阵 $[T]$，用它来对主因子载荷矩阵进行正交变换操作，最后满足式 (2.103) 为最大的条件．

K.F. Kaiser 的方差最大旋转法是现时最流行的正交旋转方案．因为它具有不管数据矩阵的大小如何而都能产生相同的数据聚类的能力．

Quartimin 法除了正交性这一点不同之外，本质上同Quartimax是相同的．在该法中，特征轴被进行斜旋转，以便使一个数据点的载荷将在一根轴上增大而在所有其他的轴上减小，这样一来，它的载荷的内积的加和就会被减小．根据Carroll 的研究，能使Quartimin函数 M 为最小即可达到上述目的．用数学公式表示，那就是使

$$M = \sum_{j<l=1}^{n} \sum_{k=1}^{c} \lambda_j \tilde{c}_{jk}^2 \lambda_l \tilde{c}_{lk}^2, \tag{2.104}$$

为最小．式中 j 和 l 分别表示第 j 个和第 l 个斜因子．

Oblimax 也是一种斜旋转方法，该法通过减少那些中间大小的载荷而增加在一指定轴上的低载荷和高载荷的数目．D.R. Saunders 指出，通过使 Kurtosis 函数，K 为最大即可达到目的．用数学公式表示，那就是使

$$K = \frac{\sum_{j=1}^{n} \sum_{k=1}^{c} \lambda_j^2 \tilde{c}_{jk}^4}{(\sum_{j=1}^{n} \sum_{k=1}^{c} \lambda_j \tilde{c}_{jk}^2)^2} \tag{2.105}$$

为最大．

方差最小 (Covarimin) 是方差最大法的扩展，它允许进行斜旋转．在这种方法中，特征向量轴作斜旋转直至方差最小函数，C 为最小．用数学公式表示，就是使

$$C = \sum_{j<l=1}^{n} [1/c \sum_{k=1}^{c} \lambda_j \tilde{c}_{jk}^2 \lambda_l \tilde{c}_{lk}^2 - 1/c^2 \sum_{k=1}^{c} \lambda_j \tilde{c}_{jk}^2 \sum_{k=1}^{c} \lambda_l \tilde{c}_{lk}^2]. \tag{2.106}$$

这些方法都已广泛地被应用于行为科学中．其中某些在自然学科，如地质学中也被应用．但是，在化学中都还没有充分的应用．

应该注意，旋转包括对旋转类型的选择，执行计算和解释结果．要求了解详细的旋转知识的化学工作者最好去参阅抽象因子分析的其它有关书刊．虽然，方差最大旋转是一种广泛地被应用的技术，但斜旋转可能对分类化学数据来说显得更加有用．对于旋转解的诠释需要仔细检验旋转所得的矩阵中的模式．要去寻找两类信息：行指定或列指定当中的聚类和对于因子的物理解释．以数据的聚类为基础的信息分类对化学工作者是非常有用的．相类似的指定在一给定的因子上具有相类似的余因子，这便表明在因子空间形成一个聚类．许多在数学上被用来比较向量的方法被用来评价指定的聚类．分散存在于因子空间的孤立区域的异常的指定则具有与任何其他的指定不相同的余因子．如果一个指定是唯一能在一给定的因子上显示出一个大的余因子的指定，则该指定对于那个因子来讲就是唯一有关系的．涉及两个或 3 个因子的旋转余因子图显示出数据的主要的聚类，采用方差最大旋转的一个三维图的例子见图 9.3.

为了发展对旋转向量的物理诠释，因子分析工作者研究了与旋转过的向量相类似的真实向量．在行为科学和其他科学研究中，已经做了许多努力试图使每一个旋转过的向量与假设的或是实验的参数等同．然而，当试图努力从物理意义上去诠释旋转过的因子时，化学工作者必须相当小心，因为期待在一个旋转过的因子与基础因子之间有一个直接的对应往往是天真的．化学工作者应该采用下一章中所介绍的目标检验而不是采用抽象旋转技术去鉴别真实因子．

2.6 数据例解

为了便于初学者对于前面所述内容在稍加熟悉之后，能有一个直观一些的数据实例来帮助加深理解和记忆，我们在这里举出一个较简单的二因子数据矩阵并对它进行抽象因子分析，列出各步骤所

获得的重要结果. 这些结果在读者自己试算时可作为对照. 对这一数据矩阵进行目标因子分析及误差干扰的讨论将分别于有关的章节中列出.

由于要举的数据矩阵是一个二因子矩阵, 故它的元素 d_{ik} 可表示成两个积项的加和

$$d_{ik} = r_{i1}c_{1k} + r_{i2}c_{2k}.$$

现随机取数分别构成行矩阵 $[R]$ 和列矩阵 $[C]$ 如下

$$[R] = \begin{matrix} a \\ b \\ c \\ d \\ e \\ f \\ g \\ h \\ i \\ j \end{matrix} \begin{bmatrix} 2 & -1 \\ 5 & 3 \\ 4 & 6 \\ 6 & 4 \\ 0 & 8 \\ 1 & 5 \\ 3 & 9 \\ 8 & -3 \\ 10 & -2 \\ 9 & 0 \end{bmatrix}, \quad [C] = \begin{matrix} \alpha & \beta & \gamma & \delta & \epsilon \end{matrix} \begin{bmatrix} 3 & 4 & 4 & 6 & 9 \\ 1 & 7 & 2 & 8 & 5 \end{bmatrix}. \tag{2.107}$$

式中, $a, b, \cdots, j$ 表示行指定, $\alpha, \beta, \cdots, \epsilon$ 表示列指定. 上面这个矩阵的相乘即可生成一个二因子矩阵 $[D]$

$$[D] = \begin{bmatrix} 5 & 1 & 6 & 4 & 13 \\ 18 & 41 & 26 & 54 & 60 \\ 18 & 58 & 28 & 72 & 66 \\ 22 & 52 & 32 & 68 & 74 \\ 8 & 56 & 16 & 64 & 40 \\ 8 & 39 & 14 & 46 & 34 \\ 18 & 75 & 30 & 90 & 72 \\ 21 & 11 & 26 & 24 & 57 \\ 28 & 26 & 36 & 44 & 80 \\ 27 & 36 & 36 & 54 & 81 \end{bmatrix}. \tag{2.108}$$

根据式 (2.19) 可以算出 $[D]$ 的协方差矩阵 $[Z]$

$$[Z]=\begin{bmatrix} 3563 & 6972 & 5012 & 9478 & 11473 \\ 6972 & 20125 & 10570 & 25410 & 24738 \\ 5012 & 10570 & 7140 & 14140 & 16408 \\ 9478 & 25410 & 14140 & 32480 & 32942 \\ 11473 & 24738 & 16408 & 32942 & 37751 \end{bmatrix}. \qquad (2.109])$$

根据下式求出 $[D]$ 的每一列的规一化常数 N_k，然后用 N_k 去除其对应的列中的各元素，便可算出 $[D]$ 的相关阵 $[Z]_N$

$$N_k=[1/\sum_i d_{ik}^2]^{1/2}, \qquad (2.110)$$

$$[Z]_N=$$

$$\begin{bmatrix} 1.00000 & 0.82334 & 0.99370 & 0.88105 & 0.98925 \\ 0.82334 & 1.00000 & 0.88178 & 0.99387 & 0.89750 \\ 0.99370 & 0.88178 & 1.00000 & 0.92852 & 0.99941 \\ 0.88105 & 0.99387 & 0.92852 & 1.00000 & 0.94076 \\ 0.98925 & 0.89750 & 0.99941 & 0.94076 & 1.00000 \end{bmatrix}. \qquad (2.111)$$

利用 JACOBI 法对 $[Z]$ 进行特征分析，得出相应的特征值 λ_1 和 λ_2 以及特征向量 V_1 和 V_2

$$\lambda_1=96790.16, \qquad \lambda_2=4268.84,$$

即

$$[\lambda^{\ddagger}]=\begin{bmatrix} 96790.16 & 0.0 \\ 0.0 & 4268.84 \end{bmatrix}, \qquad (2.112)$$

$$V_1'=(0.180783 \quad 0.440426 \quad 0.264500 \quad 0.572676 \quad 0.612719),$$

$$V_2'=(0.305979 \quad -0.562385 \quad 0.293812 \quad -0.415479 \quad 0.575459).$$

实际运算时，因为 $[Z]$ 是一个 5×5 方阵，故计算机会打印出 $\lambda_1, \lambda_2, \cdots, \lambda_5$ 等 5 个特征值和 $V_1, V_2, \cdots, V_5$ 等 5 个特征向量. 然而，后

3 目标因子分析

3.1 概　　述

60 年代发展起来的目标因子分析对检验物理模型提供了特殊的数学手段. 这一方法对化学应用显示出独特的优点.

如果我们通过理论知识、实验知识或直观感觉获得与数据矩阵中的行指定有关的物理或化学参数，用这些参数构成一个向量，它就是所谓的“目标”因子. 目标因子分析的目的就是要对一类这样的“目标”逐个单独地进行检验，以确认其是不是真实的因子. 然后，用已确认出的一组真实因子去构造数据模型，这样做的结果可经验地表示出原数据矩阵中的信息，可以做出有价值的预报.

主因子解在所有的因子分析计算中是一组基础，目标因子分析也是在这一基础上进行的. 目标因子分析的主要步骤如图 3.1 所示

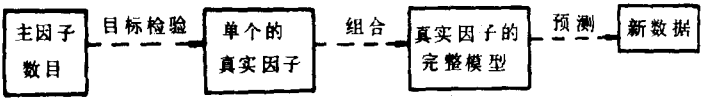


图 3.1　目标因子分析法的主要步骤框图

3.2 目标检验与目标变换

目标检验是目标因子分析的重要内容，而要对“目标”进行检验就必须对目标进行一定的变换. 目标变换在抽象解和真实解之间起着一种数学桥梁的作用. 化学工作者更加希望能得到真实的因子的模型，而目标检验和变换恰好能为此发挥更重要的作用. 在对目标

进行检验和变换时，往往涉及两种因子：典型因子和基础因子（在抽象旋转中则往往只涉及主因子和旋转了的因子）．所谓典型因子，简单地说就是可被用作因子的原始数据矩阵中的行或列．基础因子则是那些可用来描述数据矩阵中的行指定或列指定的性质的因子．通过目标检验，我们可以对与因子的性质有关的想法进行评价，进而对数据推演出具有物理意义的模型．

3.2.1 目标检验

不管问题的复杂性如何，我们都可以单独逐个地去检验潜在的因子．通过注重考虑行阵或列阵，检验只与数据矩阵的行指定或列指定有关．如果我们着重考虑行阵，则通过目标变换用以对行指定因子进行目标检验的手续可用下式来概括

$$[R]_{n,\mathrm{PFA}}T = R_{\mathrm{pred}} = R_{\mathrm{test}}, \tag{3.1}$$

式中，$[R]_{n.\mathrm{PFA}}$ 是从以 n 个因子为基础（见式 (2.87)）的主因子分析所得来的行矩阵，其余 3 个量则都是向量，T 是目标变换向量，来自一个最小二乘操作的结果，这个操作涉及到主因子分析解和各单个的由向量 R_{test} 所代表的被检验“目标”．如果被检验的向量是一个真实因子，从式 (3.1) 所得到的预测向量 R_{pred} 便与被检验的向量相当地相似，从而确认由被检验向量具体代表的概念，反之，则被检验的参数就不是一个真实的因子，从而可否认由被检验向量所代表的某些想法．

用以评价目标检验成功性的数学判据，因涉及到误差理论问题，故将放在第七章中具体讨论．

为了阐述目标检验的能力，考虑表 3.1 中所示的两个检验，它们与 10.2 节中所讨论的吸光度问题有关．通过寻找与各个纯组分的吸收光谱有关的基础因子，我们可用目标因子分析来鉴别出混合物中的组分，对于此处被检验的两个被怀疑组分，在 23 个波长处测得的（吸光系数）构成两个向量，如表中所示，可以单独进行每一个检验．

表 3.1 目标（由吸光组分在 23 个不同波长点的吸光系数构成）检验的例子

	检验 1		检验 2	
波长点序	检验向量	预测向量	检验向量	预测向量
1	0.00587	0.00598	0.00049	0.00127
2	0.00651	0.00661	0.00058	0.00130
3	0.00734	0.00747	0.00068	0.00123
4	0.00738	0.00740	0.00078	0.00110
5	0.00650	0.00654	0.00084	0.00094
6	0.00562	0.00551	0.00092	0.00080
7	0.00449	0.00440	0.00099	0.00067
8	0.00348	0.00327	0.00105	0.00053
9	0.00260	0.00240	0.00112	0.00045
10	0.00195	0.00174	0.00118	0.00036
11	0.00143	0.00120	0.00123	0.00031
12	0.00112	0.00091	0.00128	0.00029
13	0.00099	0.00076	0.00131	0.00033
14	0.00130	0.00108	0.00135	0.00046
15	0.00247	0.00238	0.00141	0.00071
16	0.00581	0.00577	0.00156	0.00109
17	0.01340	0.01333	0.00187	0.00161
18	0.02519	0.02522	0.00237	0.00218
19	0.03661	0.03651	0.00296	0.00281
20	0.04292	0.04288	0.00354	0.00351
21	0.04311	0.04306	0.00409	0.00416
22	0.03898	0.03923	0.00468	0.00479
23	0.03432	0.03428	0.00538	0.00556

由于复原步骤曾指明该问题中存在 6 个因子，所以，这两个检验中的每一个都涉及到这 6 个最重要的主因子. 通过式 (3.1) 的目标因子分析所得到的最小二乘向量在表中被列为预测向量，在第一个检验中，检验向量与预测向量之间的逐点对应性相当好，表明检验组分 1 是一个真实组分；而在第二个检验中，对应性相当差，这表明检验组分在混合物中是不存在的.

目标检验是化学数据模型化的唯一方法，这一技术具有 2 个重要特性：①每个因子可独立地被予以评价，即使是在一大批其他因

子同时影响数据时也是这样．为检验所感兴趣的向量，不必去辨别或说明其他的因子，尤其是，这一特性将目标因子分析同多元回归分析加以区别．数据的总体复杂性并不能阻碍单个的真实因子的数学分离．②通过目标组合手续可推演出真实因子组合的完整模型（这点在以后有关组合的章节中将详细加以讨论），通过目标检验，因子分析手续可用来预测新的数据．③目标检验不仅可用来确证理论，还可用来扩充和修改理论模型．当对数据缺乏内在的了解时，目标检验可用来作为一个逐项地搜寻经验模型的向导．

目标检验要求有更科学的输入．只有科学地以合理的主意为根据时，目标变换才会产生有价值的结果．事实上，如对数据没有一个预先指明的模型，则目标因子分析将是困难的，它可能会陷入一个搜索的游戏之中．所以，理论的、经验的了解和直觉等都应被用来构造检验向量和诠释目标检验的结果．

3.2.2 总体考虑

在进行目标检验之前，必须考虑两个普遍的要求：①一个检验向量所需的点的数目；②由检验值所跨越的范围．

一个检验向量中的点的数目必须大于在变换中所采用的因子的数目．有人把这一重要却又易于被忽略的判据称为大于 n 规则．例如，一个目标检验涉及到 4 个因子，则在检验向量中至少要包含有 5 个检验点，被包含进检验向量中的点越多结果就会变得越可靠．如一个检验向量中只有少于 4 个点，则这样的检验就不应该进行，因为其结果将是不可信的．如果采用 4 个点，也许能完满地预测向量，不过，这种由目标因子分析的“数学制品”所导致的肤浅的成功结果，往往会把人们引入歧途，须加以注意．

只要遵循大于 n 规则，检验点便可安全地从检验向量中被忽略．这种对于“自由浮动”的自由度是特别有用的，因为在绝大多数的检验中，许多检验值无法得到，不选用“自由浮动”方案，也许只会有非常少的化学参数能被加以检验．而且，通过自由浮动一个新的

检验向量中的少数几个已知的值可以确认目标检验的成功性，如果这些经过审慎地删除的点被充分地加以预测，那么，该检验向量的正确性就可以得到进一步的确认.

第二个对目标检验的普遍的要求关系到由检验点所跨越的范围. 每一个目标向量应尽可能完全地跨越被检验的性质的值的整个范围，对于一个检验向量，如果随便地对它进行限定，使其并不包含有那些有代表性的高值或低值，则它有可能给出与被检验值的完整范围精心限定过的检验向量完全不同的结果，实际上，应加入所有能得到的极值，希望该向量会跨越真实因子的完全范围. 一般地说，内推法较外推法更安全.

如果在对因子数目的确定时存在着许多的不确定性，则应对 $n-1$， n 和 $n+1$ 个因子进行目标检验. 随着结果的累积，也许能对因子的数目作一个较确定的估算，例如，如果 n 个基本因子的变换对 n 个因子很差而对 $n+1$ 个因子却是成功的话，那么，因子的数目可能较原来假设的要大 1 个单位. 在某些检验中，如果因子的数目大于正确的数目时，对于那些自由浮动点所做的预测值的精确度要变得小一些. 在这种情况下，因子的数目可能较原来所假设的要小一个单位.

为了获得经验的关联，各种检验向量的函数形式，诸如检验点的平方，倒数和对数等可被用来进行目标检验. 通过对检验向量的不同函数形式的检验，在目标组合步骤中会得到具有低 RMS 误差的解. 当然，采用太多的函数形式来进行检验，也会使化学工作者受到过多的缺乏物理意义的数学步骤的困扰.

3.2.3 构造检验向量

设计有化学意义的检验向量在目标因子分析中是最重要和最困难的任务.

检验向量应以理论、经验或化学见解为基础. 如果有理论的或经验的模型可利用，那么，模型中的变量支配着将被进行目标检验

的有关章节中，我们将举出许多结构的和物理的向量的例子.

表 3.2 中的 4 个假设的检验向量阐明了结构的向量的不同类型，每一个向量的基本原理在表底下予以说明.

表 3.2 结构的检验向量的例子

行指定	检验向量			
	1	2	3	4
1	1	0	2.3	22.7
2	1	0	1.6	22.4
3	0	2	-	0.3
4	-	2	5.7	-
5	0	1	3.4	0.1
6	0	1	0.8	0.5

检验对象：检验向量 1　定性标度，在该标度上，指定 1 和 2 有性质显现，指定 3, 5 和 6 缺乏性质，指定 4 被自由浮动；检验向量 2　半定量标度，在此标度上，相对于指定 5 和 6，指定 3 和 4 的性质被两倍的加权，指定 1 和 2 缺乏性质；检验向量 3　定量标度，在此标度上，指定 3 被自由浮动；检验向量 4　定量标度，在此标度上，指定 4 被自由浮动.

表中的检验向量 1，作为结构的向量的最简单的示例，是下一节中将要叙述的独特性检验的一种扩充形式. 这种“聚类”独特性检验对被认为具有被检验性质的指定用 1 表示，而对被认为缺乏被检验的性质的指定用 0 来表示，对怀疑的指定则用自由浮动值来表示. 这些“是 - 非”二元分类向量很有用，例如，可确定一个特定的官能团对一个因子是否有影响. 表中检验向量 2 用一个半定量的标度来对检验向量中的指定分等级，这样的标度对那些被认为具有性质的指定方便地采用整数值，对那些缺乏性质的指定采用 0，对不确定的或丢失的值则采用自由浮动检验点. 表 3.2 中的检验向量 3 和 4 是以详细的理论或经验知识为基础的定量向量的例子. 如果两个向量彼此互成比例，则已知具有最好精度的向量也许会更好地代表真实因子. 检验向量 1 和 4 具有相类似的模式，从目标因子分析的观点来看，彼此间实质上是等价的. 注意到检验向量 1 是一个简化了的定性标度而检验向量 4 被表达到 3 位有效数字，那么，我们

可期望检验向量 4 将会是一个真实因子的更好的代表. 通过注意那些指定在独特性检验和在旋转过的因子上产生聚类以及通过确定那些指定在典型向量的组合 (见 3.5.2 节) 中是等价的便可提示出另外的结构的向量. 变换时取得不明确的成功的检验向量可能对新的检验向量的形成提供有价值的暗示. 例如, 在一个检验向量中的差的预测点在该检验向量中可被自由浮动, 不过, 需注意, 对于为了得到越来越好的检验向量的研究不应退化成一种缺乏化学输入信号的数学练习.

3.2.4 目标变换

要进行目标检验, 首先就得对 "目标" 进行合适的变换, 进而才能对所得的预测值进行评价. 在化学问题中, 目标变换这一技术显得特别重要, 且具有很强的独特性, 因为, 尽管数据空间是复杂的, 但目标变换允许我们单个地去搜寻基础因子. 只要考察一下式 (2.98), 这一点就是显而易见的了. 该式涉及到包含在变换特征向量轴中的数学操作. 从该式可知, 新变换了的行矩阵 $[\overline{R}]$ 的第 l 列 $\overline{R}_l$ 可通过行矩阵 $[R^{\ddagger}]$ 去乘变换矩阵的第 l 列, T_l 而得到

$$\overline{R}_l = [R^{\ddagger}]T_l. \tag{3.2}$$

我们称 $\overline{R}_l$ 为预测向量, T_l 为对应变换向量. 我们希望找到该变换向量, 它能产生一个与 $\overline{\overline{R}}_l$ 有很好匹配的 $\overline{R}_l$, 这里, $\overline{\overline{R}}_l$ 是我们所怀疑的检验向量, 可能是一个基础因子, 也也正是我们在前面所多次提到过的 "目标". 要做到这一点, 我们选用最小二乘手续, 它使得检验向量和预测向量之间的偏差为最小, 这一手续对于被考虑的单个的目标检验会产生最佳可能的变换向量, 我们将在此详细地描述获得最佳 T_l 的数学基本原理.

变换向量 T_l 含有元素 $t_{1l}, t_{2l}, \cdots, t_{nl}$, $[R^{\ddagger}]$ 的每一行可被当作一个行指定向量. $[R^{\ddagger}]$ 的第 i 行是一个向量 R'_i, 它含元素 r_{i1}, r_{i2}, $\cdots$, r_{in}. 向量 R'_i 不应与 R_i 相混淆, R_i 是行因子矩阵的第 i 列. 当 R'_i 被 T_l 相乘, 可得到 $\overline{r}_{il}$, 即第 i 个实体在新的目标变换了

的坐标轴上的投影

$$\bar{r}_{il} = R_i' T_l = r_{i1}t_{1l} + r_{i2}t_{2l} + \cdots + r_{in}t_{nl}, \tag{3.3}$$

式 (3.3) 的加和复盖全部 n 个因子.

用 T_l 去乘行矩阵的每一个行向量便得出 $\bar{r}_{1l}, \bar{r}_{2l}, \cdots, \bar{r}_{rl}$, 它们就是预测向量 $\bar{R}_l$ 的元素. 然后，将预测向量的每一个元素与检验向量 $\bar{\bar{R}}_l$ （含元素 $\bar{\bar{r}}_{1l}$， $\bar{\bar{r}}_{2l}, \cdots, \bar{\bar{r}}_{rl}$ ）相比较. $\bar{r}_{il}$ 与 $\bar{\bar{r}}_{il}$ 之间的差为 Δr_{il}

$$\Delta r_{il} = \bar{r}_{il} - \bar{\bar{r}}_{il} = r_{i1}t_{1l} + r_{i2}t_{2l} + \cdots + r_{in}t_{nl} - \bar{\bar{r}}_{il}. \tag{3.4}$$

为找到最佳的 T_l ，通过设由式 (3.4) 所示的差的平方的导数加和等于零而使检验向量与预测向量之间的偏差为最小. 例如，差平方对 t_{1l} 的导数为

$$\frac{\mathrm{d}(\Delta r_{il})^2}{\mathrm{d}t_{1l}} = 2r_{i1}^2 t_{1l} + 2r_{i1}r_{i2}t_{2l} + \cdots + 2r_{i1}r_{in}t_{nl} - 2r_{i1}\bar{\bar{r}}_{il}, \tag{3.5}$$

对于每一个行指定得到相类似的表达式，对所有的行指定加和并应用最小二乘判据，得到

$$\begin{aligned} \sum_{i=1}^{r} \frac{\mathrm{d}(\Delta r_{1l})^2}{\mathrm{d}t_{1l}} &= 0 \\ &= t_{1l}\sum_i r_{i1}^2 + t_{2l}\sum_i r_{i1}r_{i2} + \cdots \\ &\quad + t_{nl}\sum_i r_{i1}r_{in} - \sum_i r_{i1}\bar{\bar{r}}_{il}, \end{aligned} \tag{3.6}$$

重复上述计算，对于 T_l 的其余的元素 $t_{2l}, t_{3l}, \cdots, t_{nl}$ ，我们使偏差平方的加和为最小. 于是，得到下列方程组

$$\left.\begin{aligned} \sum r_{i1}\bar{\bar{r}}_{il} &= t_{1l}\sum r_{i1}^2 + t_{2l}\sum r_{i1}r_{i2} + \cdots + t_{nl}\sum r_{i1}r_{in}, \\ \sum r_{i2}\bar{\bar{r}}_{il} &= t_{1l}\sum r_{i1}r_{i2} + t_{2l}\sum r_{i2}^2 + \cdots + t_{nl}\sum r_{i2}r_{in}, \\ &\cdots\cdots\cdots\cdots \\ \sum r_{in}\bar{\bar{r}}_{il} &= t_{1l}\sum r_{i1}t_{in} + t_{2l}\sum r_{i2}r_{in} + \cdots + t_{nl}\sum r_{in}^2. \end{aligned}\right\} \tag{3.7}$$

为了用矩阵符号来表达这些方程式，定义两个向量

$$A_l = \begin{bmatrix} \sum r_{i1}\bar{\bar{r}}_{il} \\ \sum r_{i2}\bar{\bar{r}}_{il} \\ \cdots \\ \sum r_{in}\bar{\bar{r}}_{il} \end{bmatrix} \quad \text{和} \quad T_l = \begin{bmatrix} t_{1l} \\ t_{2l} \\ \cdots \\ t_n l \end{bmatrix}, \tag{3.8}$$

以及下面的矩阵

$$[B] = \begin{bmatrix} \sum r_{i1}^2 & \sum r_{i1}r_{i2} & \cdots & \sum r_{i1}r_{in} \\ \sum r_{i1}r_{i2} & \sum r_{i2}^2 & \cdots & \sum r_{i2}r_{in} \\ \cdots & \cdots & \cdots & \cdots \\ \sum r_{i1}r_{in} & \sum r_{i2}r_{in} & \cdots & \sum r_{in}^2 \end{bmatrix}. \tag{3.9}$$

用矩阵符号表示，式 (3.7) 现在可写成

$$A_l = [B]T_l, \tag{3.10}$$

用 $[B]^{-1}$ 乘以两边，得

$$T_l = [B]^{-1}A_l. \tag{3.11}$$

从式 (3.9) 和 (2.83)，可以见到

$$[B] = [R^{\ddagger}]^T[R^{\ddagger}] = [\lambda^{\ddagger}], \tag{3.12}$$

式中，$[\lambda^{\ddagger}]$ 是一个仅由基本特征值构成的对角矩阵. 再考虑式 (3.8)，可得出结论

$$A_l = [R^{\ddagger}]^T\bar{\bar{R}}_l, \tag{3.13}$$

式中，$\bar{\bar{R}}_l$ 是检验向量，它由与行指定有联系的有猜疑的参数组成. 因此

$$T_l = [\lambda^{\ddagger}]^{-1}[R^{\ddagger}]^T\bar{\bar{R}}_l. \tag{3.14}$$

这个方程式是目标因子分析的中心方程. 作为变换矩阵 $[T]$ 的一个列，T_l 是最小二乘向量变换器，它很容易由此方程算得.

有了式 (3.14)，对目标因子分析工作者来说，目标变换变得容易了. 除了检验向量 $\overline{\overline{R}}_l$ 外，该方程中所有其他的数值都在常规的分解步骤中自然获得. 当然，如前面章节所述，检验向量必须根据理论知识、经验知识或是直觉去构造，因此，可以一点也不夸张地说，如何推演出检验向量已构成了包含在目标因子分析中的真正的化学技术. 为了弄明白某一个被猜疑的因子是不是一个真实因子，可将这一检验向量代入式 (3.14) 中，这时就可得到最佳的可能变换向量 T_l.

得到 T_l 后，运用式 (3.2) 便可获得 $\overline{R}_l$ 的各元素值. 然后，我们便可确认下述方程是否在实验误差范围内成立

$$\overline{R}_l \stackrel{?}{=} \overline{\overline{R}}_l. \tag{3.15}$$

如果我们所怀疑的检验向量 $\overline{\overline{R}}_l$ 是一个因子，则 $\overline{R}_l$ 的每一个元素将在实验误差范围内与 $\overline{\overline{R}}_l$ 的对应元素相等；反之，各对应元素之间的差值将大于我们的期待值. 对于这种比较的细节将在 3.4 节中做一般的描述. 某些较定量的判断检验向量可信性的方法将于第七章中详细阐述.

3.3 自由浮动与迭代目标检验

在构造检验向量时，由于数据占有的不充分或对某些行指定 (或列指定) 性质的猜疑，致使所构造出来的检验向量 (即我们所要检验的“目标”) 不一定是完整的. 事实上，不必一定要求检验向量是完整的. 对于检验向量中的一些未知点 (也称为空白数据点)，在检验时可让其留下空白，随意用 0 或其他数值暂时顶替，然后采取某些适当的手续，在进行目标检验的同时，又对这些空白点予以预测. 有两种手续可以完成这样的任务，一种称为自由浮动，另一种称为迭代目标检验.

3.3.1 自由浮动

通过目标检验，可以对数据做常规的预测. 检验向量不必一定

是完整的，检验点的未知值可留下空白，这就是一种称为自由浮动的手续．鉴于目标检验的特点，预测值总是完整的．在一个成功的目标检验中，被自由浮动的点自动地被予以预测．这样一来，尚未被量测的基础数据可根据目标因子分析而得到预测，例如，在表 3.1 中的第一个目标检验．如果该被检验化合物在某一波长处的克分子吸光系数不知道，则那一检验点在目标检验中可以被自由浮动．既然该检验向量是一个真实因子，那么，那一检验点的值将会被精确地予以预测．

从数学上讲，3.2.4 节所述的用以进行变换的最小二乘法完全是一般性的．既便是某一特定的检验向量中的某些 $\overline{r}_{il}$ 值是空白时，也是可行的．不过这时，需从式 (3.7) 到 (3.11) 的加和中除去某些恰当的项，然后必须用式 (3.11) 来计算向量 T_l，因为式 (3.14) 已不再有效．通过这些适当的修改后，可用式 (3.2) 来计算 $\overline{R}_l$ 并将所得结果与实验性的检验向量 $\overline{\overline{R}}_l$ 相比较．这种叫做自由浮动的手续具有一种潜在的优点，方程式 (3.2) 对每一个指定，包括那些在检验向量中被自由浮动的，都自动地产生一个 $\overline{r}_{il}$ 值．

对目标检验向量有一个重要的约束．在检验向量中检验点的数目必须大于 n (即数据矩阵的秩)．当然，输入一个检验点数目不足的检验向量总是会对检验点数产生完全的吻合的，不过，这样的结果总是毫无意义的，所以，必须注意，在采用目标检验手续时，所用检验向量必须包含有大于因子数目的检验点．关于这方面的详细讨论将于 7.4 节中给出．

为阐述自由浮动，让我们回到图 2.5 去．该图显示了一个给定数据矩阵的 5 个列（5 种性质）之间的向量联系．采用最小二乘目标变换技术，我们试图去寻找一个具有物理意义并处在这一平面上的参考轴，如果偶极矩是一个真实因子，那么，一个行指定（代表物质的一个分子）点在偶极矩轴上的投影就产生出该分子的偶极矩．为确定这一个轴的位置，我们不需知道所有行指定分子的偶极矩．对于一个二因子空间，在数学上，要求知道的偶极矩的最小数目为

2, 然而，建议采用多于两个的检验点.

实际上，自由浮动所有其偶极矩未知的行指定分子也可构造一个偶极矩向量. 如前所述，只用数据点的不完全集来计算变换向量，作为式 (3.2) 的结果，计算机对全部的分子都输出其偶数极矩值，不管这些分子在目标检验流程中是否被涉及. 这样便可以对偶极矩进行预测，虽然，这并不是因子分析的目的，但对于化学工作者来说，这的确是一个重要的收获.

3.3.2 迭代目标检验

前面已经讲过，只要某检验数据集复盖性质的因子空间，自由浮动这种数学目标检验手续可被用来对检验向量中的空白点提供预测. 这里再介绍另一种涉及到迭代技巧的目标检验手续，它同样可用来对目标数据中的空白点进行预测.

回想式 (3.2) 和 (3.14) 分别为

$$\overline{R}_l = [R^{\ddagger}]T_l,$$

$$T_l = [\lambda^{\ddagger}]^{-1}[R^{\ddagger}]^T\overline{\overline{R}}_l,$$

从上两式，可得

$$\overline{R}_l = [R^{\ddagger}][\lambda^{\ddagger}]^{-1}[R^{\ddagger}]^T\overline{\overline{R}}_l. \tag{3.16}$$

用一个普通的矩阵 $[P]$ 代替上式中用以变换 $\overline{\overline{R}}_l$ 成 $\overline{R}_l$ 的各矩阵乘积，得

$$\overline{R}_l = [P]\overline{\overline{R}}_l. \tag{3.17}$$

为简便，将式 (3.17) 变成

$$\overline{R} = [P]\overline{\overline{R}}, \tag{3.18}$$

$$[P] = [R^{\ddagger}][\lambda^{\ddagger}]^{-1}[R^{\ddagger}]^T = \begin{bmatrix} A & B & \cdots & C \\ D & E & \cdots & F \\ \cdots & \cdots & \cdots & \cdots \\ G & \cdots & H & I \end{bmatrix}. \tag{3.19}$$

$$[R^{\ddagger}] = \begin{bmatrix} 7.069 & 0.182323 \\ 6.90975 & 0.8712755 \\ 9.074753 & -0.805546 \end{bmatrix}.$$

设检验向量 $\overline{\overline{R}}$ 为：(1.00 0.50 0.00)，预测值列于表 3.3 中.

表 3.3　目标检验结果

检验向量	预测向量		
	1	2	3
1.00	0.49	1.000	1.001
0.50	0.777	0.499	0.499
0.00	0.185	2.003	2.005

预测向量 1　不进行迭代的目标检验结果；预测向量 2　进行迭代的目标检验结果；预测向量 3　此时，将检验向量中的空白值定为 100.0，并进行迭代目标检验.

由表 3.3 可见，预测向量 1 的结果由于在检验向量中引入 0 而被歪曲了，预测向量 3 的结果证明迭代目标检验法并不受空白点的初始值设置的影响（设置 0.0 和 100.0，最后所得结果一样）.

事实上，数据矩阵 $[D]$ 是对应于 x 分别为 1.0，2.0 和 3.0 时，下列 3 条直线的值

$$y_1 = x + 2, \quad y_2 = 0.5x + 3, \quad y_3 = 2.0x + 1.$$

所以，它的真实的行因子和列因子应为

$$[D] = \begin{bmatrix} 3.0 & 4.0 & 5.0 \\ 3.5 & 4.0 & 4.5 \\ 3.0 & 5.0 & 7.0 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 0.5 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \end{bmatrix}.$$

这一手续的优点是可以直接通过矩阵 $[P]$ 来对检验向量中的空白点提供预测值.

A. Lorber 等以投影矩阵的基本性质为基础概括了目标因子分析和迭代目标因子分析的数学的和统计学的性质. 他们的概括提醒化

学工作者应该尽量学会采用精确的现代数学公式来描述自己的数据分析方法.

3.4 目标检验结果的诠释

诠释和评价目标检验的结果是一项困难的任务. 有两种完整的途径, 一种以数学判据为基础, 另一种以定性比较为基础, 被用来进行评价目标检验的结果. 由于实验误差以复杂形式影响目标检验, 定量的判据迄今尚未被充分地加以评估, 定性的途径可能会引入歧途. 故在实践中, 对目标检验结果的最终评价应以来自这两种途径的综合结论为依据.

一种以实验误差对目标检验的影响为依据的数学途径 (见 7.5 节) 特别有用, 该法的一个主要优点是考虑了检验点中的误差和数据矩阵中的误差这两者的影响, 应用两个函数, SPOIL 和 RELI, 来评价目标检验的结果.

还有许多其他的用于评价两个向量的相似性的数学的和统计学的判据. R.J. Rummel 在他的著作中得出结论: RMS 系数、叠合系数和组合相关系数对于比较向量特别有用, 读者感兴趣可去参阅他的著作. 对于目标检验, D.G. Howery 及其合作者推荐组合相关系数, 而 R.W. Rozett 等则采用相对因子误差, 此外, 统计学 F 检验和特征值的频率分析等方法也将在第七章中加以讨论. 因此, 对与目标检验有关的各种数学判据进行比较研究是很有必要的.

为了对目标检验做定性评价, 就必须用化学的眼光去对检验向量和它的对应的预测向量仔细地进行逐点的比较. 为阐述定性途径, 表 3.4 和 3.5 列出几个假设的结果, 表 3.4 中的例子包含完整的检验向量, 表 3.5 中的则包含带有自由浮动点的检验向量.

目标检验一个完整的检验向量的 5 个假设的结果列于表 3.4 中, 检验向量中点的误差被定为小于 5, 以检验向量与预测向量对于这一误差的定性比较为依据所得的结论被列于表的底下. 在第三个结果

表 3.4 目标检验一个完整的检向量的假设结果

行指定	检验	假设的预测结果				
	向量	1	2	3	4	5
1	238	281	292	244	279	147
2	134	138	131	111	230	273
3	42	44	59	73	32	114
4	187	180	177	235	188	215
5	255	258	250	216	261	122
6	91	85	83	123	94	156

定性结论（假设有两个因子，假设检验点中的误差小于 5）：结果 1 非常好的相似性，检验向量是一个基础因子；结果 2 好的相似性，检验向量是一个基础因子；结果 3 一般的相似性，检验向量是一个模棱两可的基础因子；结果 4 好的相似性，但指定 2 的预测值差，检验向量不是一个基础因子；结果 5 差的相似性，检验向量不是一个基础因子.

表 3.5 目标检验一个不完整的检验向量的假设结果

行指定	检验	假设的预测结果			
	向量	1	2	3	4
1	114	119	132	55	97
2	63	66	56	79	83
3	-	98	94	112	24
4	-	57	61	-2	-36
5	88	80	81	103	73
6	-	73	103	4	179

定性结论（假设有两个因子，假定检验点中的误差小于 5）：结果 1 非常好的相似性，检验向量是一个基础因子，对被自由浮动的点的预测值也许是可信的；结果 2 一般的相似性，检验向量是一个模棱两可的基础因子，对被自由浮动的点的预测值也许是可信的；结果 3 差的相似性，检验向量不是一个基础因子，对被自由浮动点的预测是不可信的；结果 4 一般的相似性，虽然对指定 4 和 6 的预测值有疑问地超出检验点的范围，但检验向量也许是一个模棱两可的基础因子，对自由浮动点的预测值也许会是不可信的.

中阐述的这一类结论在目标因子分析中是普遍的，常规工作中不应被抛弃，在评价这种模棱两可的结果时，对检验向量性质的化学了解与来自上面讨论过的定量途径的结论的综合考虑应该起着最终的

作用. 虽然结果 4 指出一个不成功的检验，但这种结果可能会提出令人感兴趣的新的检验向量，在这种情况下，其预测值较差的指定 2 的检验值可以被自由浮动以构成一个新的目标向量.

由于数据的不充足或了解的不够充分，在一个检验向量中往往有几个点不得不留下空白. 这样一种不完整的检验向量在化学问题中是典型的. 表 3.5 列出了一个包含有被自由浮动点的检验向量所产生的 4 种假设结果，由于假设有两个因子（即 $n=2$）. 且在检验向量中已给出 3 个点，这就遵循了大于 n 规则，因此，检验本身应被认为是靠得住的，以检验向量与预测之间的相似性为依据的定性结论列在表的底下.

对这类涉及到被自由浮动点的目标检验，必须予以特别的注意. 因为随着被自由浮动的点的数目的增加，一个检验向量会显出更差的定义性，当然，这是以化学工作者对来自应用自由浮动点的检验的结论具有较小的自信心为条件而言的. 预测向量 4 含有两个可疑值：一个对于指定 6 来说是相对大的值和一个对于指定 4 的负值. 一旦预测值超出输入检验点的范围，就得小心谨慎. 如果检验向量被拙劣地设置在一个完整的基础因子的高的或是低的端值时，上述情形就会发生. 除非能从化学的观点上使不正常的预测值合理化，否则，即便变换看起来是成功的，但该因子的物理意义仍应被认为是有问题的.

3.5 组合及有关的预测

目标因子分析的目的是用已通过目标检验确认出的一个真实因子集来构造数据模型，从数据矩阵中抽提出更有用的信息并做出有价值的预测，进而使人们了解那些对数据产生物理意义的基本的影响. 通过组合步骤，我们可以发现真实因子的一个“关键集”，也就是说获得最佳的真实因子模型，并且以其为依据去预测新的数据行和数据列.

3.5.1 组合概述

发现足够数目可接受的检验向量是可能的，但这依然决定不了因子空间，因为某些检验向量分布在一个共同的子空间中，因此，必须通过目标组合步骤来对真实因子的完整模型加以检验．通过对不同的真实因子集的检验结果做比较，才可能对某一化学问题找到最佳的目标因子分析解．

在目标组合步骤中，复原数据阵用的是真实因子而不是抽象因子．行阵 $[R]_{\text{real}}$ 是由经选择过的一套 n 个真实因子所组成的．当然，这些因子都已事先成功地通过了目标检验．为进行这样的组合，我们采用目标变换矩阵 $[T]$，这一矩阵是由 n 个目标变换向量构成的，每一个变换向量都是用 $[R]_{\text{real}}$ 中的每一个对应的真实因子通过式 (3.14) 而得到的．用 $[T]^{-1}$ 代表 $[T]$ 的逆，按式 (2.97)，用它来变换 $[C]_{\text{PFA}}$ 成为一个新的列矩阵 $[C]_{\text{comb.}}$ 组合．这样，目标组合手续可用下面的方程式和比较来加以概括

$$\begin{aligned}[R]_{\text{real}}\{[T]^{-1}[C]_{\text{PFA}}\} &= [R]_{\text{real}}[C]_{\text{comb.}} \\ &= [D]_{\text{comb.}} \stackrel{?}{=} [D], \end{aligned} \tag{3.24}$$

如果 $[R]_{\text{real}}$ 代表问题中所有的因子，那么，组合 - 复原矩阵 $[D]_{\text{comb.}}$ 将合理地与原数据矩阵 $[D]$ 相似．因此便可确证被检验的这个真实因子的组合的可信性．式 (3.24) 实质上可改写成

$$[\overline{R}][\overline{C}] = [D^{\ddagger}]_{\text{TFA}} \stackrel{?}{=} [D], \tag{3.25}$$

$[D^{\ddagger}]_{\text{TFA}}$ 是目标组合复原矩阵．如果它与数据阵 $[D]$ 在实验误差范围内匹配，则可因此而确信已找到一个合适的真实因子集，这就意味着目标组合检验获得成功．对于成功的组合检验，$[R]_{\text{real}}$ 中的因子被叫做关键因子．有时，通过理论原理可以确定一个最好的真实因子集，要不然，也可通过检验真实因子的所有的或是遴选过的组合（每个组合包含 n 个因子）来确定真实因子的一个最佳集，能以

此，任何两个数据列均可被用来确定一个行指定点，然后，可通过读出在其余数据列向量上的投影来进行预测，通过最小二乘方法，重新确定参考轴使其能与原数据矩阵中的一套合适的典型列相一致．采用原数据的一个列作为检验向量，即可做到这一点，因为每一个列都在因子空间中，故每一个列都会产生一个成功的检验，检验手续将会产生一个变换向量，这样的变换向量的一个组合需要用来构造变换矩阵．一般地说，数据列的一个任意组合将未必能产生一个能够成功地复原数据的变换矩阵，只有能旋转总的因子空间的某一些组合才能完成上述任务，能最佳地完成上述任务的组合集被叫做关键组合集．

现在，让我们来简介一下上述过程所涉及的数学步骤．数据矩阵的第 l 列可被当作向量 $\overline{\overline{D}}_l$，它包含有该列中的数据点，这些数据点代表该向量的各元素，换言之，$\overline{\overline{D}}_l$ 被用作检验向量，令 $\overline{\overline{D}}_l = \overline{\overline{R}}_l$，应用式 (3.14)，我们定义

$$T_l = [\lambda]^{-1}[R]^T\overline{\overline{D}}_l, \tag{3.27}$$

在组合中，用 n 个这样的变换向量构造一个完整的变换矩阵

$$[T] = [T_a \ \ T_b \ \ \cdots T_n]. \tag{3.28}$$

在努力寻找能最好地复原数据矩阵的关键组合集的过程中，各种组合均被应用．因此，我们寻找一套 n 个数据列，它能够最好地满足

$$[\overline{\overline{D}}]_{\text{key}}[\overline{C}] - [D] = \text{最小}, \tag{3.29}$$

式中，

$$[\overline{\overline{D}}]_{\text{key}} = [\overline{\overline{D}}_a \ \ \overline{\overline{D}}_b \ \ \cdots \ \ \overline{\overline{D}}_n], \tag{3.30}$$

和

$$[\overline{C}] = [T]^{-1}_{\text{key}}[C^{\ddagger}]. \tag{3.31}$$

在这些表达式中，有下标“ Key ”的表示关键组合集．

利用数据列做为空间的因子有许多优点，首先，它们较从因子分析所得到的抽象因子更易于想象，其次，它们的应用排除了鉴别真正起控制作用的因子的必要．对于许多化学问题，这已经是足够的了．这一技术的足有成效的应用在以后的章节中都可见到．

为了进一步的阐述应用在关键组合集中的原理，让我们再回到图 2.4 和 2.5 所描绘的例子中去，在那里，所研究的是分布在同一平面中的 5 个数据列向量．因为所有 5 个列向量都分布在同一平面上，所以，它们是线性相关的．图 2.4 中 5 个向量中的任何两个都可用来做为代表轴．例如，我们选择 C_1 和 C_5 为基轴，如图 3.2 所示，从 C_2,C_3 和 C_4 的顶端画平行于 C_1 和 C_5 的线，每一基轴上相交的长度代表基轴向量在 C_2, C_3 和 C_4 这 3 个向量上的载荷．从图 3.2 推出 5 个向量可表达如下

$$\left.\begin{array}{rcrcl} C_1 &=& 1.0000C_1 &+& 0.0000C_5, \\ C_2 &=& 1.0000C_1 &+& 1.4124C_5, \\ C_3 &=& 1.3986C_1 &+& 0.8313C_5, \\ C_4 &=& 1.2122C_1 &+& 1.3722C_5, \\ C_5 &=& 0.0000C_1 &+& 1.0000C_5. \end{array}\right\} \tag{3.32}$$

这些线性方程式有一个重要的特点，一个行指定点在 C_2, C_3 和 C_4 上的投影可由在基轴 C_1 和 C_5 上的得分来计算．这样一来，由一个大的数目的数据列构成的一个数据集可被变成具有代表性的数据列的一个最小的集．从它出发，所有的数据都可被生成．通过应用这些方程式，我们不再需要计算在从因子分析结果所得到的抽象特征向量上的投影，而只需贮存那些与所选择的代表轴 C_1 和 C_5 有关的数据．

应用合适的计算机程序，化学工作者可以检验典型向量的选择性组合或是原数据矩阵中向量的全部可能的组合．如果研究人员只希望检验为数不多的想法或是手头所拥有的计算能力有限的话，那么，检验选择性组合就可以了，检验全部可能的组合当然可以保证

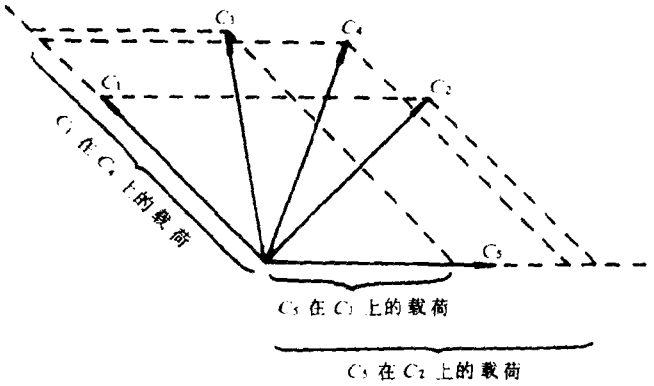


图 3.2　在代表轴 C_1 和 C_5 上的因子载荷

找到典型向量的最佳集，然而，却要花费大量的计算机机时. 数据矩阵的大小和每次组合所取的向量数目决定了组合的总次数. 例如，如果这种组合步骤涉及一个含有 4 个因子解的 12×14 矩阵，那么，典型列的组合的数目便是 $14 \times 13 \times 12 \times 11/4! = 1001$，设列数为 c，因子数为 n, 则组合数目就是 $c!/((c-n)!n!)$, 即 $14!/((14-4)!4!)$, 对于大的矩阵，计算机用于执行一个完全的组合步骤的时间很容易就会超过 1 小时. 如果只有较小的计算机，那就应在一个缩小了的规模上来执行组合步骤. 当然，一个被缩小了的组合步骤应该包括有这样的典型向量，它们是化学研究人员最感兴趣的，它们复盖性质的较广的范围，且具有大的独特性值（见 3.6 节）.

通过下述淘汰步骤可有效地鉴别关键典型向量. 首先，运行一个经化学了解而选择出的数目较小的组合以确定一个初步的关键集，然后，每次用一个其他的向量去取代第一套组合中的一个向量以发现一个得到改善的第二个关键集，取代过程重复到所需要的次数，每一阶段都采用前一阶段所获得的最佳集. 这样做，从相对来说为数不多的组合可能会产生出一个有效的解.

这样一来，第三个关键行将是使上面所示的行列式的值离零最远的行．采用类似的方式，可从归一化后的行因子矩阵中提取出每一个连续的关键行，一般地说，第 m 个关键行应该是使

$$\det \begin{vmatrix} v_{i1} & v_{i2} & \cdots & v_{im} \\ v_{j1} & v_{j2} & \cdots & v_{jm} \\ \cdots & \cdots & \cdots & \cdots \\ v_{m1} & v_{m2} & \cdots & v_{mm} \end{vmatrix}$$

离零最远的行．将原数据矩阵转置，然后采用上述的同样手续可获得典型的关键集．

在组合中采用典型向量可得到几种信息．每个组合模型的成功与否可通过比较该模型的 RMS 误差与 RMS 实验误差而加以判断，具有明确的因子分析解的化学问题可被期望有相当好的典型向量解．一般地说，典型向量关键组合集的 RMS 误差只稍微大于采用相同数目的因子的抽象复原的 RMS 误差．用一小套典型向量可以较好地代表因子，对一些选择过的较好的组合，通过对其所涉及的指定和 RMS 误差列表便可以归纳其结果，在最佳的组合中最常被显示的指定可以同那些只从化学意义上被挑选出的指定相媲美．

从组合步骤得到的结果对于分类指定是有用的．例如，可以用这两种方式来鉴别相类似的指定：如果两种不同的组合导致近似相等的 RMS 误差，则这两套向量是等价的，尤其是，如果一个典型向量被另一个典型向量取代而在组合检验中并不产生有意义的 RMS 误差变化，则这两个互相交换的典型向量是等价的，这两个指定属于相同的聚类，通过从许多组合中比较等价性，便可检测出聚类的总的模式．为鉴别最重要的典型向量，将每一向量出现在最佳组合中的次数的频率加以列表将会是很有价值的，在最佳模型中出现频率最大的典型向量最重要．

应用上面目标组合步骤获得的典型向量关键集可以预测数据矩阵的新数据行和新数据列，因而也就能扩展数据矩阵．如为了预测一个与新的行指定 x 和一个原来的列指定 k 有关的数据，我们采用

因子分析的基本方程的一个变更形式

$$d_{rk}(\text{预测}) = \sum_{j=1}^{n} r_{rj}(\text{已知}) \cdot C_{jk}(\text{组合}), \tag{3.33}$$

列余因子 C_{jk} 取自根据式 (3.24) 计算得到的矩阵 $[C]_{\text{comb.}}$，但是行余因子的值必须从与因子分析无关的来源得到. 采用典型因子的关键集，我们当然可以对 $[C]_{\text{comb.}}$ 进行计算.

为了演示计算，考虑一个服从 2 因子模型 4×5 数据矩阵，希望对某些新的行指定 (用 $a, b, \cdots, m$ 来标记) 预测完整的数据行.

首先，应用目标组合步骤可找到数据矩阵中可给出最佳数据复原的一对典型列，假设例子的关键集包含原数据阵中的列 2 和列 3，那么，我们可计算出相对应的已变换过的列矩阵

	列指定				
关键因子	1	2	3	4	5
2 (与关键的列 2 对应)	C_{21}	1	0	C_{24}	C_{25}
3 (与关键的列 3 对应)	C_{31}	0	1	C_{34}	C_{35}

(3.34)

这里，余因子 C_{jk} 是用关键列 2 和列 3 的组合计算而得到的. 接着，我们便可形成一个包含新的行指定的新的列矩阵

$$\begin{array}{cc} \text{新行指定} & \text{关键因子} \\ & \begin{array}{cc} 2 & 3 \end{array} \\ \begin{array}{c} a \\ b \\ \cdots \\ m \end{array} & \begin{bmatrix} d_{a2} & d_{a3} \\ d_{b2} & d_{b3} \\ \cdots & \cdots \\ d_{m2} & d_{m3} \end{bmatrix} \end{array} \tag{3.35}$$

这里，余因子 d_{ij} 是与新的行指定和关键的典型列 (即原数据阵的列 2 和列 3) 有关的关键数据，对于任何给定的行指定，两个关键余因子 d_{ij} 必须是已知，用行矩阵 (式 (3.3 5)) 前乘列矩阵 (式 (3.34))，便可生成一个完整的目标因子分析 - 预测数据矩阵. 例如 d_{b4}，它

与新的行指定 b 和非关键列指定 4 有关，那么，可通过用矩阵（式 (3.34)）的第四列去同矩阵（式 (3.35)）的第 b 行逐项相乘而获得

$$d_{b2}C_{24} + d_{b3}C_{34} = d_{b4}\ (\text{被预测的}), \tag{3.36}$$

式 (3.36) 与因子分析的基本方程式具有相同的形式.

上面所述的方法，具体算例将于本章的最后一节中提供.

由于鉴定典型向量的关键集不需对数据有任何事先的了解，如有必要，做为例行工作，目标预测步骤可被常规地进行. 假定关键集对原来的数据是一个好的模型，如果新的指定象原来的指定一样服从相同的因子模型，则以典型向量为基础的预测被期望是可信赖的. 当然，如能对某些被预测的点得到实验值，则预测的可信性便可直接得到证明.

3.5.3 基础向量的组合及有关的预测

前已述过，所谓基础向量就是描述数据矩阵中的行指定或列指定的性质的因子. 基础向量使人们了解对数据产生影响的基本物理或化学意义.

通过检验基础因子的组合，我们试图以一种基础的方式来构造数据模型，即寻找下式的最后解

$$[R]_{\text{basic}}[C]_{\text{basic}} = [D], \tag{3.37}$$

这里，$[R]_{\text{basic}}$ 和 $[C]_{\text{basic}}$ 中的每一个因子都是基础因子. 方程式 (3.37) 表达了目标因子分析的最高目的. 这类型的解是因子分析的完美结果. 虽然构造基础因子需要做很大的努力，但是，目标因子分析能够产生用其他方法所不能得到的详细模型.

我们可以通过在 3.2.4 节中所描述的那些手续，对基础向量单个地去进行检验和辨别. 由于表达真实化学因子方式的多样性及由于化学因子间的多方面的内部联系，所以，发现能充分地旋转因子空间的基础因子的一个“关键集”确实不是一件容易的事情. 为找到这一关键集，可以接受的基础因子的各种组合就形成行因子矩阵

$[\overline{\overline{R}}]_{\text{basic}}$，并进行下面的计算

$$[\overline{\overline{R}}]_{\text{basic}}[\overline{C}]_{\text{basic}} = [D^{\ddagger}]_{\text{basic}}, \tag{3.38}$$

式中

$$[\overline{C}]_{\text{basic}} = [T]^{-1}[C^{\ddagger}]. \tag{3.39}$$

这里，用变换矩阵 $[T]$ 的逆 $[T]^{-1}$ 来左乘抽象的列因子矩阵而得到基础列因子矩阵 $[\overline{C}]_{\text{basic}}$，而变换矩阵 $[T]$ 是由与基础因子有关的各单个的变换向量构成的. 如果组合复原数据 $[D^{\ddagger}]_{\text{basic}}$ 充分地等于原始的数据矩阵

$$[D^{\ddagger}]_{\text{basic}} \approx [D], \tag{3.40}$$

则可认为基础因子的一个关键集已找到.

目标 - 组合步骤可被应用于已经选择过的基础向量或是全部已成功地被变换过的基础向量，计算设备和研究的目的对于决定基础向量的组合范围起着支配的作用.

如果组合检验要包罗所有的基础向量，不管其在目标检验中是属于模棱两可的抑或是较成功的，那么，组合的总的数目可能是会大得惊人的. 例如，若有 25 个向量通过目标检验，在一个 6 因子问题中，将会有 25!/((25-6)!6!)=25 × 24 × 23 × 22 × 21 × 20/6! =177100 次组合. 完成这样的工作，即便是在较大的计算机上也需要耗费大量的执行时间. 实际上并不需要执行全部可能的组合，因为许多的基础向量对可能在物理上是等价的. 此外，应用一些科学判据可以减小在组合中被检验向量的数目至一个易于处理的程度，如具有最低的 SPOIL 值（见 7.5.3 节）的向量对于组合中的向量来说是特别好的侯选者. 另方面，我们可以淘汰那些经变换后确认是完全不成功的向量以及那些在采用化学的或目标因子分析的判据时显示出等价的向量.

在组合中应采用完整的向量. 如果在组合中采用那些具有自由浮动过的点的向量，复原矩阵将缺少对应于每一个被自由浮动的指定的元素，为解决这一问题，对于每一个空白点，将插入来自预测

向量的相应的值，这时，组合将产生一个完整的矩阵，不过，须注意，在组合中如采用较多的预测值，则组合所得结果的可信赖性就会变得较小，如果是由于疏忽而在组合中只用预测向量而不是用检验向量，则结果所得的解将只是一个缺乏真正输入信息的抽象解，对这样的解理应剔除.

研究人员应努力通过理论和经验的知识去判断关键的组合解，如果所进行的因子分析研究是以合理的理性模型为依据且经过仔细的筹划，则组合步骤就会进行得比较顺利. 用化学理解去关联组合-目标因子分析模型是目标因子分析的最满意的最终产物. 由于绝大多数化学问题的复杂性，基础向量的关键组合集不一定非要在实验误差范围内复原数据. 一般地说，在一个困难的问题中，一个目标因子分析模型，如果其 RMS 误差为实验误差的 3 倍，则它可能应被认为是非常令人满意的了.

组合的结果可被用来鉴别等价的基础向量，也可用来发现在较好的目标因子分析模型中那些最常被描述的基础向量，对于进行理论模型与经验模型的比较以及对于为数据确定实际模型来说，这些信息都是有价值的. 至于用来提取这类信息的方法已于 3.5.2 节中做过解释.

运用基础向量关键集在一个扩展数据中预测新的点的步骤与在 3.5.2 节所描述的步骤类似. 当然，只有在目标组合中已经鉴别出较好的基础向量集之后，才能开始这一步骤.

现在，以一套 n 个基础向量来开始讨论. 假定该基础向量集在组合步骤中已满足要求地复原了数据. 为预测用 a, b, $\cdots$, m 标识的新的行指定有关的数据，对每一个新的指定，每一个关键的基础余因子的值都是需要的. 例如，假定关键组合集在一个 2 因子问题中已被显示出包含有两个用 x 和 y 标识的关键参数，那么，我们可以采用在 3.5.2 节中所描述过的步骤. 在这里，列矩阵将具有下面的形式

关键因子	列指定					
	1	2	3	4	5	
x（与关键参数 x 有关）	C_{x1}	C_{x2}	C_{x3}	C_{x4}	C_{x5}	(3.41)
y（与关键参数 y 有关）	C_{y1}	C_{y2}	C_{y3}	C_{y4}	C_{y5}	

式中，余因子 C_{jk} 通过包含有关键参数 x 和 y 的目标组合来进行计算. 包含有新的行指定的新行矩阵为

$$\begin{array}{cc} \text{新行指定} & \text{关键因子} \\ & \begin{array}{cc} 2 & 3 \end{array} \\ \begin{array}{c} a \\ b \\ \cdots \\ m \end{array} & \begin{bmatrix} d_{ax} & d_{ay} \\ d_{bx} & d_{by} \\ \cdots & \cdots \\ d_{mx} & d_{my} \end{bmatrix} \end{array} \tag{3.42}$$

式中，行余因子 d_{ij} 是新的行指定的关键参数的已知值. 用行矩阵（式 3.42）左乘列矩阵（式 3.41）就可以产生目标因子分析预测数据的一个矩阵. 例如，新的数据 d_{c5} 可通过下式给出

$$d_{cx}c_{x5} + d_{cy}c_{y5} = d_{c5}(\text{预测的}), \tag{3.43}$$

对从基础向量的一个组合预测得到的数据，可期望其有这样一个 RMS 误差，它最好也只不过等于、也许会大于当采用基础向量的那一关键集去复原数据阵时所得到的 RMS 误差. 如果关键因子不旋转新的指定的因子空间，则预测数据 RMS 误差往往相当的大. 如果一个新的指定同某些原来的行指定在化学上相似，则对该指定的预测更可能是可信的. 只要是有可能，应该通过同量测值加以比较来检验目标预测值.

3.6 独特性检验和单位向量检验

在目标因子分析研究中，在一般情况下，应对行或列指定常规地进行一种被称为独特性检验的特殊目标检验，这种检验被设计用来鉴别或是因化学的独特性或是因数据中的严重误差而引起具有不

1 和第 6 个指定具有中等大的预测值，分别为 0.37 和 0.42, 可初步确认这两个指定属同一聚类. 来自一个线性自由能问题的独特性检验请参阅第九章.

独特性检验的结果常常用列表方式来表示. 对每一个指定，给出在它的检验中的独特性值，列出每一次检验中其预测值大于 0.15 的其他指定，后一信息可用来将相似的指定集合入聚类中，对一个色谱问题的独特性概括载入表 9.7 中，在因子的一定范围内对每一个指定的独特性值进行列表对于将独特行为同特定因子相联系是有价值的. 例如，当因子大小从 $J-1$ 增大至 J 时，如果某一个指定的独特性值发生戏剧性的增大，则该指定的某些独特性质可以说明第 J 个因子.

另一种应常规进行的特殊检验是单位向量检验，这种检验向量完全由 1 这个数构成，是用来检验一个对所有指定来说都是共同的常数因子的. 如果从目标变换所得的预测值都接近等于 1，则单位 1 就被鉴别为一个因子，如果所有的行指定都与相同的官能团相结合，则可期望单位检验向量有一个好的变换. 如果第一个主因子特别重要，则不管物理状况如何，单位检验向量也将有好的变换. 取自部分化学问题所得的单位向量检验列于第九章中.

3.7 数据例解

目标因子分析是以抽象因子分析的结果为基础的. 在 2.6 节中，我们已对式 (2.108) 所示的数据矩阵 $[D]$ 完成了抽象因子分析，得到了由式 (2.113) 和 (2.114) 所示的抽象结果. 那么，如何通过目标变换来从 $[R^{\dagger}]$ 和 $[C^{\dagger}]$ 进而求得真正的结果 $[\overline{R}]$ 和 $[\overline{C}]$ 呢？本节将演示解答这一问题的各个步骤.

在 3.2.3 节中，我们已经阐述过，设计有化学或物理意义的检验向量 (即目标) 在目标因子分析中是最重要、最困难的任务. 这种设计应以理论、经验或化学直觉为基础. 在此，假设我们已设计出或

怀疑有以下 3 个检验目标，分别用 $\overline{\overline{R}}_1$, $\overline{\overline{R}}_2$ 和 $\overline{\overline{R}}_3$ 来表示

$$\left.\begin{array}{rcl}(\overline{\overline{R}}_1)' &=& \begin{array}{ccccc}(\ 2 & 5 & 4 & 6 & 0\\ 1 & 3 & 8 & 10 & 9\ ),\end{array}\\ (\overline{\overline{R}}_2)' &=& \begin{array}{ccccc}(-1 & 3 & 6 & 4 & 8\\ 5 & 9 & -3 & -2 & 0\ ),\end{array}\\ (\overline{\overline{R}}_3)' &=& \begin{array}{ccccc}(\ 3.8028 & 2.8704 & 1.5008 & 3.5072 & 5.5417\\ 7.8648 & 3.7760 & 2.4177 & 6.8098 & 2.4107\ ).\end{array}\end{array}\right\} \quad (3.43)$$

现在，问题归结到它们到底是不是真正的目标？为了解答这一问题，可做以下计算. 首先根据式 (3.14) 分别算出它们的变换向量 T_1, T_2, 和 T_3

$$T_l = [\lambda^{\ddagger}]^{-1}[R^{\ddagger}]^T\overline{\overline{R}}_l,$$

式中的 $[\lambda^{\ddagger}]$ 和 $[R^{\ddagger}]$ 分别如式 (2.112) 和 (2.114) 所示，则可得

$$\begin{aligned} T_1 &= \begin{bmatrix} 96790.16 & 0.0 \\ 0.0 & 4268.84 \end{bmatrix} \\ &\quad \begin{bmatrix} 13.1874 & 95.8762 & \cdots & 110.8132 \\ 8.5494 & 2.1806 & \cdots & 22.7691 \end{bmatrix} \overline{\overline{R}}_1 \\ &= \begin{bmatrix} 0.0485326 \\ 0.1590727 \end{bmatrix}, \end{aligned}$$

同样可以算得

$$T_2 = \begin{bmatrix} 0.0351851 \\ 0.1712394 \end{bmatrix} \quad 和 \quad T_3 = \begin{bmatrix} 1.6640610 \\ -2.9117990 \end{bmatrix}.$$

再根据式 (3.2)，便可分别算出上述 3 个检验目标的预测值 $\overline{R}_1$,

$\overline{R}_2$ 和 $\overline{R}_3$

$$\overline{R}_1 = \begin{pmatrix} 2.0000 \\ 5.0000 \\ 4.0000 \\ 6.0000 \\ 2.86 \times 10^{-6} \\ 1.0000 \\ 3.0000 \\ 8.0000 \\ 10.0000 \\ 9.0000 \end{pmatrix}, \quad \overline{R}_2 = \begin{pmatrix} -1.0000 \\ 3.0000 \\ 6.0000 \\ 4.0000 \\ 8.0000 \\ 5.0000 \\ 9.0000 \\ -3.0000 \\ -2.0000 \\ 7.15 \times 10^{-7} \end{pmatrix}, \quad \overline{R}_3 = \begin{pmatrix} 1.76791 \\ 2.37220 \\ 3.13869 \\ 3.96126 \\ 4.83140 \\ 5.40952 \\ 5.72166 \\ 4.94552 \\ 4.44145 \\ 2.90579 \end{pmatrix}.$$

将 $\overline{R}_1, \overline{R}_2$ 和 $\overline{R}_3$ 分别同 $\overline{\overline{R}}_1, \overline{\overline{R}}_2$ 和 $\overline{\overline{R}}_3$ 作比较，可以见到，前两个目标的预测值同检验值有十分好的匹配，故可以断定它们都是真正的因子，而 $\overline{R}_3$ 同 $\overline{\overline{R}}_3$ 的值的却完全大不相同，因此，可断定它不是一个真实因子. 这样，$\overline{R}_1$ 和 $\overline{R}_2$ 的组合即可得到变换后新坐标系下的行矩阵

$$[\overline{R}] = [\overline{R}_1 \ \ \overline{R}_2], \tag{3.44}$$

T_1 和 T_2 的组合即可得到变换矩阵

$$[T] = \begin{bmatrix} 0.0485326 & 0.0351851 \\ 0.1590727 & -0.1712394 \end{bmatrix}.$$

根据式 (2.101)，可算出新坐标系下的列矩阵 $[\overline{C}]$

$$\begin{aligned} [\overline{C}] &= [T]^{-1}[C^{\ddagger}] \\ &= \begin{bmatrix} 0.0485326 & 0.0351851 \\ 0.1590727 & -0.1712394 \end{bmatrix}^{-1} \\ &\quad \begin{bmatrix} 0.180783 & \cdots & 0.612719 \\ 0.305979 & \cdots & 0.575459 \end{bmatrix} \\ &= \begin{bmatrix} 3.000001 & 4.000003 & 3.999999 & 6.000001 & 9.000006 \\ 1.000000 & 6.999998 & 1.999999 & 7.999996 & 4.99999 \end{bmatrix} \end{aligned} \tag{3.45}$$

式 (3.44), (3.45) 所示的 $[\overline{R}]$ 和 $[\overline{C}]$ 与式 (2.107) 所示的完全吻合. 可见 $[\overline{R}]$ 和 $[\overline{C}]$ 便是我们所要求的真正的行阵和列阵. $[\overline{R}]$ 和 $[\overline{C}]$ 之所以同式 (2.107) 所示的值有如此好的吻合, 那是因为矩阵 $[D]$ (式 (2.108)) 是由式 (2.107) 所示的 $[R]$ 和 $[C]$ 经矩阵乘法准确获得的. 可以说, $[D]$ 是一个不带误差的 "纯" 矩阵. 此外, 所设计的两个检验目标也是 "纯" 的目标. 如果 $[D]$ 中渗入误差, 那情况又会怎样呢? 我们将在 7.7 节中讨论和演示这一问题.

下面我们演示一下如何通过典型向量组合来进行数据的预测. 假设我们取式 (2.108) 所示的原始数据阵 $[D]$ 中的第一列和第二列分别为典型向量 $\overline{\overline{D}}_1$ 和 $\overline{\overline{D}}_2$, 根据式 (3.27) 可求出由它们各自的变换向量构成的变换矩阵

$$
\begin{aligned}
[T]_{\text{key}} &= [\lambda^{\ddagger}]^{-1}[R^{\ddagger}]^{T}[\overline{\overline{D}}_1 \quad \overline{\overline{D}}_2] \\
&= \begin{bmatrix} 96790.16 & 0.0 \\ 0.0 & 4628.84 \end{bmatrix} \\
&\quad \begin{bmatrix} 13.1874 & 95.8762 & \cdots & 110.8132 \\ 8.5494 & 2.1806 & \cdots & 22.7691 \end{bmatrix} \begin{bmatrix} 5 & 1 \\ 18 & 41 \\ \cdots & \cdots \\ 27 & 36 \end{bmatrix} \\
&= \begin{bmatrix} 0.1808 & 0.4404 \\ 0.3060 & -0.5624 \end{bmatrix}.
\end{aligned}
$$

再根据式 (3.31) 计算经变换过的列因子矩阵为

$$
\begin{aligned}
[\overline{C}] &= [T]_{\text{key}}^{-1}[C^{\ddagger}] \\
&= \begin{bmatrix} 2.3786 & 1.8628 \\ 1.2942 & -0.7646 \end{bmatrix} \\
&\quad \begin{bmatrix} 0.180783 & 0.440400 & 0.264500 & 0.572676 & 0.612719 \\ 0.305979 & -0.562385 & 0.293812 & -0.415479 & 0.575459 \end{bmatrix}
\end{aligned}
$$

$$= \begin{bmatrix} 1.0000 & 0.0000 & 1.1765 & 0.5882 & 2.5294 \\ 0.0000 & 1.0000 & 0.1176 & 1.0588 & 0.3529 \end{bmatrix}.$$

如果我们要预测式 (2.108) 所示矩阵中的元素 d_{34} 和 d_{93} ，则可参照式 (3.36) 计算得

$$\begin{aligned} d_{34} &= d_{31}\times c_{14} + d_{32}\times c_{24} \\ &= 18\times 0.5882 + 58\times 1.0588 \\ &= 72.0, \end{aligned}$$

$$\begin{aligned} d_{93} &= d_{91}\times c_{13} + d_{92}\times c_{23} \\ &= 28\times 1.1765 + 26\times 0.1176 \\ &= 36.0 \end{aligned}$$

4. 秩消因子分析

4.1 问题的提出与基本思想

在许多实际工作中，分析工作者往往会碰到这样的分析任务：即在待分析的混合物中仅对某些组分含量的确定是有意义的，而其余的组分的确定是无关紧要的. 在这种情况下，如果有一种方法在不考虑样品中分析工作者不感兴趣的其他成分的情况下能获得所要分析的已知组分的定量信息，那么这种方法将是理想的. 经典的最小二乘法是一种概念简单且易于实现的计算方法，但是，对于多组分定量分析，最小二乘法只有对所有的主要成分是已知的体系才能得到可靠的结果. 曲线分析法也是一种可以考虑的方法之一，不过运用该法时，在数据矩阵中必须存在这样一些点，在这些点处，仅有一个组分有贡献，此外，此法也只能处理二组分混合物. 经典的因子分析方法对于解决上述问题也无能为力. 二次双线性探测器 (该仪器给出形如 $M_{ij} = \sum \beta_k X_{ik} y_{jk}$ 的二维数据) 提供的数据适合于解决上述问题，而能较好地解析二维数据矩阵的是一种叫做秩消因子分析 (亦称消秩因子分析，简称 RAFA) 的技术，该法已被成功地应用于萤光混合物分析以及混合物体系的液相色谱分析.

秩消因子分析可简略地描述如下 (为易于理解，这里以萤光仪产生的激发－发射光谱为例): 假设一个多组分溶液的激发－发射光谱矩阵 (Excitation-Emission Matrix 以下称为 EEM) 为 $[M]$，其秩数在理论上应等于溶液中存在的萤光物种数. 如果我们知道其 EEM 为 $[N]$ 的某一组分存在于该溶液中，并且假使我们从 $[M]$ 中减去 $[N]$ 的确切的量，则 $[M]$ 原来的秩应减小 1，在这种情况下，我们将观察到，$[M]$ 中对应于 $[N]$ 的特征值变为零. 由于实际的实验数据中

存在着误差，其特征值不可能完全的消失，然而，它将会达到一个极小．在对应的特征值达到极小时，被减去的 $[N]$ 的量与混合物中该已知成分的相对浓度相对应，具体地说，我们知道组分 k 存在于一多组分的溶液中，对于其他共存的组分或者了解或者不了解．为了对组分 k 进行定量，在相同的实验条件下对多组分溶液和浓度已知的一个 k 组分标准溶液获得它们各自的 EEM，设其分别为 $[M]$ 和 $[N]$．设多组分溶液中存在的组分 k 的浓度为 $c(k)$，组分 k 的标准溶液的浓度为 $c^0(k)$，两者的比值为 $\beta(k)=c(k)/c^0(k)$．如果从矩阵 $[M]$ 减去 $\beta(k)[N]$，则结果所得的修正矩阵的秩将比 $[M]$ 小 1，即此时 $[M]$ 中对应于 $[N]$ 的特征值变为零．由于存在实验误差，故该特征值永远达不到零而只能达到一个极小．通过不断变换标量 β 的值，并从 $[M]$ 减去 $\beta[N]$ 并求取结果所得的修正矩阵的秩，当修正矩阵的秩比 $[M]$ 小 1 时的 β 值就是我们所要寻找的 $\beta(k)$ 值．知道了 $\beta(k)$ 值和 $c^0(k)$ 值，就可求解 $c(k)$ 的值．这就是秩消因子分析方法的基本思想．

4.2 秩消因子分析的早期模型

对于单组分，其 EEM 的理想形式为

$$M_{ij}=\alpha X_i Y_j, \tag{4.1}$$

这里，X_i 与在波长为 λ_i 处发射的光子数目 (即发射强度) 成正比，Y_j 与在波长 λ_j 处吸收的光子数目 (即激发强度) 成正比．如果 X_i 和 Y_j 是规一化的 (即 $\sum X_i^2=\sum Y_j^2=1$) 以及在所有的波长处的吸收都是低的，那么数 α 就与 X_i 和 Y_j 无关并且正比于浓度．对于一个由 r 个组分组成的混合物，在理想条件下

$$[M]=\sum[M^{(k)}]. \tag{4.2}$$

这里 $[M^{(k)}]$ 是假设只有组分 k 存在时得到的矩阵．如方程 (4.1) 和 (4.2) 满足，则 $[M]$ 可写成

且

$$[N^{(k)}] = \alpha_k^0 X_k Y_k^T. \tag{4.12}$$

则 α_k^0, X_k 和 Y_k 分别与 ξ_1, U_1 和 V_1 完全相等，用标准矩阵 $[N^{(k)}]$ 去解式 (4.5)—(4.10) 可知道这一点.

在建立秩消方法之前，让我们考虑一下标准的最小二乘方法. 如果所有的成分都已知，那么，我们可以定义一误差矩阵 $[E]$

$$[E] = [M] - \sum_{k=1}^{r} \beta_k [N^{(k)}], \tag{4.13}$$

总的平方误差 ϵ 定义为

$$\epsilon = \sum E_{ij}^2 = \mathrm{Tr}[E]^T[E] = \mathrm{Tr}[E][E]^T \geq 0. \tag{4.14}$$

这里，$\mathrm{Tr}[A]$ 代表 $[A]$ 的迹. 使 ϵ 最小化将给出 $\beta_k = c_k/c_k^0$ 和 $\epsilon = 0$. 如果所有的组分都是未知的，最小二乘将缺乏合理性，如

$$[E] = [M] - \sum_{k=1}^{p} \beta_k [N^{(k)}] - \sum_{k=r+1}^{q} \beta_k [N^{(k)}], \tag{4.15}$$

此处 $p < r$ 仅是某些确实存在的组分，而 $k > r$ 是某些想象中存在但实际上不存在的组分，这时 $\epsilon = \mathrm{Tr}[E]^T[E]$ 在用 $\beta_k = c_k/c_k^0$ 估算时不一定是最小，在这种情况下，使 ϵ 最小化不能作为寻找 β_k 的合理手续.

如果一个特定的待测组分确切地存在于混合物中，则向量集 X 和 Y 就应分别处在由 $\{U_k\}_{k=1}^r$ 和 $\{V_k\}_{k=1}^r$ 各自所组成的向量空间中，这时，由式 (4.13) 所定义的 $[E]$ 在 $\beta(k) = c/c^0(k)$ 时，它的秩就会变成 $r-1$. 这里，c 是混合物中被测定的组分的浓度，$c^0(k)$ 是纯组分 k 的标准溶液的浓度. 而且，具有特征值 $\mu_1 \geq \mu_2 \geq \cdots \geq 0$ 的矩阵 $[E]^T[E]$ 和 $[E][E]^T$ 将会给出一个 $\mu_r(\beta)$，它在 $\beta = \beta(k) = c/c^0(k)$ 处有最小值 (由于实验误差的存在，这一最小值不会达到零)，因此，$\mu_r(\beta)$ 的极小能定义出正确的 $\beta(k)$.

全体. 我们将 $\sigma_i = \sqrt{\lambda_i}(i = 1, 2, \cdots, r)$ 称为 $[A]$ 的奇异值，等式 $[A] = [U][D][V]^T$ 则被称为 $[A]$ 的奇异值分解.

通过以前我们所介绍过的抽象因子分析法可以确定一个 $m \times n$ 的实验数据阵 $[A]$ 的 r 值. 一旦确定了 r 值，则从 r 个有意义的特征值和特征向量，应用式

$$[\overline{A}] = [\overline{U}][\overline{S}][\overline{V}]^T, \tag{4.18}$$

即可得到 $[A]$ 阵的近似阵 $[\overline{A}]$. 此时，$[\overline{U}]$ 是一个 $m \times r$ 阵，它包含有代表 $[A][A]^T$ 空间的最前面的 r 个特征向量；$[\overline{V}]$ 是一个 $r \times n$ 阵，它包含有代表 $[A]^T[A]$ 空间的最前面的 r 个特征向量；$[\overline{S}]$ 是一个包含有最前面 r 个奇异值的对角阵.

通俗地讲，上述这种一个实验矩阵或其近似矩阵分解成 3 个矩阵的过程就叫做奇异值分解 (简称 SVD). 奇异值分解在因子分析中也是一个重要的计算过程， R.I. Shrager 在他的工作中介绍了应用 SVD 进行光谱分析时的方法. 在秩消因子分析法中，借助这种算法可以对 $\beta(k)$ 值进行直接计算.

4.3.2 非迭代求解

在秩消因子分析中，对于每一种组分，它的二维数据矩阵 $[N_k]$ 应该服从这样的规则，即 $[N_k] = X_k Y_k^T$ ，这里 X_k 是一个列向量，它的元素为 $x_{ki}(i = 1, 2, \cdots, m)$ ， Y_k 是一个列向量，它的元素为 $y_{kj}(j = 1, 2, \cdots, n)$，下标 k 代表第 k 个组分. 上标 T 表示矩阵的转置操作. 在化学问题中，应在上面的方程式的右边乘上一个表达量子作用率，克分子吸光系数和浓度的标量. 为方便起见，这一标量合并入 X_k 和 Y_k 中去. 矩阵 $[N_k]$ 当作一个浓度为 c_k^0 的标准矩阵.

由 r 个组分 $(k = 1, 2, \cdots, r)$ 组成的某些混合物的数据阵则可表示成全部组分的数据矩阵的线性加和

$$[D] = \sum_{k=1}^{r} X_k Y_k^T / \lambda_k, \tag{4.19}$$

式中标量 $\lambda_k = c^0(k)/c(k)$, $c^0(k)$ 代表组分 k 在标准溶液中的浓度, $c(k)$ 代表组分 k 在该混合物中的浓度.

秩消因子分析法的巧妙之处在于观察到如果能找到 λ_k，则矩阵 $[M] = [N_k] - \lambda_k[D]$ 是一个 $(r-1)$ 秩矩阵. 这一观察提示人们，只要检查在变换 λ_k 时对矩阵 $[M]$ 的特征值所产生的影响便可对组分 k 的浓度得出一个定量的结果.

在传统的数学处理中，减小上述矩阵的秩的问题就相当于设该矩阵的行列式为零. 于是，问题可用公式表达成

$$|[N_k] - \lambda_k[D]| = 0, \tag{4.20}$$

这实际上也就是广义的特征值 / 特征向量问题

$$[N_k]Z = \lambda_k[D]Z, \tag{4.21}$$

式中，Z 是特征向量，λ_k 是特征值. 不过，这里所讨论的情况同传统的特征值问题相比较，有几个不同的地方：①矩阵 $[N_k]$ 和 $[D]$ 很可能是矩形阵；②矩阵 $[N_k]$ 是奇异的；③矩阵 $[D]$ 和 $[D]^T[D]$ 不是单位矩阵. 通过奇异值分解可帮助求解这一问题.

采用矩阵 $[D]$ 的奇异值分解，上述问题可重新表示成

$$X_k Y_k^T Z = \lambda_k[\overline{D}]Z = \lambda_k[\overline{U}][\overline{S}][\overline{V}]^T Z, \tag{4.22}$$

式中，$[\overline{D}]$, $[\overline{U}]$, $[\overline{S}]$ 和 $[\overline{V}]$ 的定义参阅 4.3.1 节所述.

运用矩阵 $[\overline{U}]$ 的正交性，得出

$$[\overline{U}]^T X_k Y_k^T Z = \lambda_k[\overline{S}][\overline{V}]^T Z, \tag{4.23}$$

定义 $Z' = [\overline{S}][\overline{V}]^T Z$，上式可写成

$$[\overline{U}]^T X_k Y_k^T [\overline{V}][\overline{S}]^{-1} Z' = \lambda_k Z', \tag{4.24}$$

上式也可写成

$$[\overline{U}]^T [N_k][\overline{V}][\overline{S}]^{-1} Z' = \lambda_k Z', \tag{4.25}$$

由于矩阵 $([\overline{U}]^T[N_k][\overline{V}][\overline{S}]^{-1})$ 是方阵，所以，这就是通常的特征值 - 特征向量问题. 因为矩阵 $([\overline{U}]^T[N_k][\overline{V}][\overline{S}]^{-1})$ 不是对称阵，所以，向

量 Z' 不是垂直的，因为 $[N_k]$ 的秩为 1，对于特征值 λ_k 将存在 $(r-1)$ 个零解. 所以，仅存的非零解将等于矩阵 $([\overline{U}]^T[N_k][\overline{V}][\overline{S}]^{-1})$ 的迹：即 $\lambda_k = \mathrm{Tr}([\overline{U}]^T[N_k][\overline{V}][\overline{S}]^{-1})$.

知道了 λ_k，则混合物中组分 k 的浓度 c_k 可由下式直接计算

$$c_k = c_k^0/\lambda_k.$$

4.3.3 秩消因子分析法与标准加入法的结合

从传统的意义上讲，对化学物种的定量存在着两种途径：标准曲线和标准加入法. 当样品中所含被研究的信息量很少时，常常选择标准加入法，有人利用一维数据通过多元线性回归手段来使传统的标准加入法通用化. 然而，正如所有的一维方法一样，这种通用的标准加入法也是有局限的，要在分析之前需要知道样品中的全部组分. 单就这一点来讲，这种方法对复杂样品的应用性就缺乏通用性. 此外，这种标准加入法同传统的标准加入法一样存在着内在的缺点. 分析探测器的响应必须予以调零. 这里所指的所谓分析探测器指的是能对某一种有分析价值的性质提供测量的实体，它是“分析信号”的来源. 这种分析信号可以通过数学变换形式对样品分析提供有用信息.

根据秩消因子分析法及标准加入法的原理，可以预料，这两种方法的结合不但可以允许在事先对其他组分完全未知的情况下测定目标组分，而且不需要“调零”. 因此，通过标准加入法使二维数据与标准相结合可对复杂样品的分析提供一种强有力的手段. 这种相结合的定量过程包括两个步骤：①获取矩阵 $[N_k]$，即从原样品的数据中扣除向其中加入已知量 c_k^0 的待测组分的样品的数据；②应用前面所讲过的秩消因子分析法中的

$$\lambda_k = \mathrm{Tr}([\overline{U}]^T[N_k][\overline{V}][\overline{S}]^{-1}) \quad 和 \quad c_k = c_k^0/\lambda_k$$

两个方程式去测定样品中组分 k 的含量.

4.4 广义秩消因子分析

前面介绍的秩消方法在计算过程中需要许多矩阵的对角化处理，而且每分析一种组分需要一校准矩阵，这在实际应用中有诸多不便. 继秩消之后，有人提出了广义的秩消因子分析法 (简称 GRAFA)，该法能确定未知混合物中各个待分析组分的双线性光谱及其相对含量. 其基本原理是：任一双线性数据矩阵 $[M]$ 可表示为 n 个纯组分双线性光谱 μ_k 的线性加和

$$[M] = \sum_{k=1}^{n} \beta_k \mu_k, \tag{4.26}$$

此处，$\mu_k = X_k Y_k^T$，且 $(\mu_{ij})_k = x_{ik} y_{jk}$，$X_k$ 是一些具有第一种信息 (如激发光谱) 的列向量，Y_k^T 是一些具有第二种信息 (如发射光谱) 的行向量. 如果定义 μ_k 为“单元浓度”，即纯组分的双线性光谱，β_k 为 $[M]$ 中第 k 种化合物的浓度，则可将式 (4.26) 写成矩阵形式

$$[M] = [X][\beta][Y]^T, \tag{4.27}$$

这里，$[X]$ 的列就是 n 个 X_k 向量，$[Y]^T$ 的行是 n 个 Y_k^T 向量，$[\beta]$ 是一个对角矩阵，其对角元素是相应的组分的浓度 β_k. 通常可获取两个数据矩阵，即未知浓度的数据阵 $[M]$ 以及标准数据矩阵 $[N]$. 双线性标准数据矩阵 $[N]$ 也可用类似的矩阵形式来表达

$$[N] = [X][\xi][Y]^T, \tag{4.28}$$

其中矩阵 $[X]$ 和 $[Y]^T$ 的定义与在方程 (4.27) 中的定义相同，$[\xi]$ 是一个对角矩阵，其对角元素是已知组分的浓度 ξ_k.

矩阵 $[M]$ 和 $[N]$ 具有相同的 $[X]$ 和 $[Y]^T$ 部分，例如激发光谱和发射光谱皆相同，仅是浓度矩阵 $[\beta]$ 和 $[\xi]$ 不同. 因此，为解出方程

在进行奇异值分解之后，根据抽象因子分析法估算合适的因子数，在理想状况下所估算的因子数目应等于存在于样品混合物中的组分数目 n, 通过 $[U]$, $[V]$ 的前 n 个“重要的”列及 n 个大的 s 值构成矩阵 $[\overline{M}]$，它在误差范围内包含有 $[M]$ 中的关键性信息

$$[\overline{M}] = [\overline{U}][\overline{S}][\overline{V}]^T, \tag{4.38}$$

现在，方程 (4.32) 可写成

$$[N][Z][\beta] = [\overline{M}][Z][\xi] = [\overline{U}][\overline{S}][\overline{V}]^T[Z][\xi], \tag{4.39}$$

若我们将 $[Z]$ 表示为 $[Z] = [\overline{V}][\overline{S}]^{-1}[Z]^*$，这里 $[Z]^* \equiv [\overline{S}][\overline{V}]^T[Z]$, 则

$$[N]([\overline{V}][\overline{S}]^{-1}[Z]^*)[\beta] = [\overline{U}][\overline{S}][\overline{V}]^T([\overline{V}][\overline{S}]^{-1}[Z]^*)[\xi]. \tag{4.40}$$

因 $[\overline{V}]$ 是正交阵，所以有 $[\overline{V}]^T[\overline{V}] = [I]$ (其中 $[I]$ 为单位矩阵), 因此

$$[N][\overline{V}][\overline{S}]^{-1}[Z]^*[\beta] = [\overline{U}]([\overline{S}][\overline{S}]^{-1})[Z]^*[\xi], \tag{4.41}$$

进一步简化为

$$([N][\overline{V}][\overline{S}]^{-1})[Z]^*[\beta] = [\overline{U}][Z]^*[\xi], \tag{4.42}$$

上式左乘 $[\overline{U}]^T$，右乘 $[\beta]^{-1}$，得

$$\begin{aligned}([\overline{U}]^T[N][\overline{V}][\overline{S}]^{-1})[Z]^*[\beta][\beta]^{-1} &= ([\overline{U}]^T[\overline{U}])[Z]^*[\xi][\beta]^{-1} \\ &= [Z]^*[\lambda],\end{aligned} \tag{4.43}$$

其中 $[\lambda] \equiv [\xi][\beta]^{-1}$，式 (4.43) 也可写成

$$([\overline{U}]^T[N][\overline{V}][\overline{S}]^{-1})[Z]^* = [Z]^*[\lambda], \tag{4.44}$$

因矩阵 $([\overline{U}]^T[N]_k[\overline{V}][\overline{S}]^{-1})$ 是方阵，故 (4.44) 式就是通常的特征值 - 特征向量方程. 由于矩阵 $([\overline{U}]^T[N]_k[\overline{V}][\overline{S}]^{-1})$ 不是对称的，所以 $[Z]^*$ 表示的特征向量不是正交的. 又由于 $[N]$ 的秩为 1，所以将会有 $p-1$

个特征值 λ_k 为零，因此，只有非零解才等于矩阵 $([\overline{U}]^T[N]_k[\overline{V}][\overline{S}]^{-1})$ 的迹. 通过计算这一矩阵的迹，即 λ_k，第 k 个组分的浓度 β_k 可直接由 $\beta_k = \xi_k/\lambda_k$ 求得.

4.4.2 几种组分的同时定量

在这种情况下，标准数矩阵 $[N]$ 含有几种存在于样品矩阵 $[M]$ 中的组分. 首先必须验证 $[N]$ 中的组分是否构成样品矩阵 $[M]$ 所含物种数的一个子集，这可通过对 $[N]$ 进行双线性目标因子分析来验证，即考察下式是否成立

$$[\overline{U}][\overline{U}]^T[N][\overline{V}][\overline{V}]^T = [N], \tag{4.45}$$

若投影矩阵 $[\overline{U}][\overline{U}]^T$ 和 $[\overline{V}][\overline{V}]^T$ 使 $[N]$ 不变，则 $[N]$ 中所含的组分就是 $[M]$ 的子集.

如果在标准矩阵 $[N]$ 中存在 1 种以上的物种，则方程 (4.44) 就有几个较大的非零特征值. 利用特征向量集 $[Z]^*$ 可计算纯组分光谱矩阵 $[X]$ 和 $[Y]^T$(如激发和发射光谱)

$$[Z]^* = [\overline{S}][\overline{V}]^T[Z] = [\overline{S}][\overline{V}]^T([Y]^T)^+, \tag{4.46}$$

$$[Y]^T = ([\overline{V}][\overline{S}]^{-1}[Z]^*)^+, \tag{4.47}$$

由定义 $[M] = [X][\beta][Y]^T = [\overline{U}][\overline{S}][\overline{V}]^T$ 得

$$\begin{aligned}[X][\beta] &= [M]([Y]^T)^+ \\ &= [\overline{U}][\overline{S}][\overline{V}]^T[\overline{V}][\overline{S}]^{-1}[Z]^* \\ &= [\overline{U}][Z]^*. \end{aligned} \tag{4.48}$$

对于每一组分来说，λ_k 是浓度比值 ξ_k/β_k (即校正 / 未知)，一旦获得了纯谱 X_k 或 Y_k^T，就容易找到那一个 ξ_k 对应于那一个比值 λ_k，因此浓度 β_k 就能够通过 $\beta_k = \xi_k/\lambda_k$ 求得.

4.4.3 以校准作为基础

在这种情况下，样品数据矩阵 $[M]$ 是校准矩阵 $[N]$ 中组分的一个子集，这时 $[M]$ 的主成分不能构成表示 $[N]$ 的一个基础，因

此，方程 (4.44) 在这种情况下不能成立. $[N]$ 的主成分可由 $[N]=[\overline{U}]_N[\overline{S}]_N[\overline{V}]_N^T$ 估算，可得到与 (4.44) 式，(4.47) 和 (4.48) 式类似的下列诸式

$$([\overline{U}]_N^T[M][\overline{V}]_N[\overline{S}]_N^{-1})[Z]_N^*=[Z]_N^*[\lambda]_N, \tag{4.49}$$

$$[Y]^T=([\overline{V}]_N[\overline{S}]_N^{-1}[Z]_N^*)^+, \tag{4.50}$$

$$[X][\beta]=[\overline{U}]_N[Z]_N^*. \tag{4.51}$$

特征值 $(\lambda_N)_k$ 的定义与前面所定义的不同，对于每一组分来说，此时 $(\lambda_N)_k=\beta_k/\xi_k$ (即，未知样 / 校正).

可用双线性目标因子分析检验 $[M]$ 中的物种是否构成 $[N]$ 的子集，$[M]$ 在由 $[N]$ 所定义的空间中的投影应使其保持不变，即

$$[\overline{U}]_N[\overline{U}]_N^T[M][\overline{V}]_N[\overline{V}]_N^T=[M], \tag{4.52}$$

这种情况也可用主成分回归或多元线性回归来求解，因为所有组分的光谱都是已知的.

4.4.4 通用模型

当遇到在校准样品中含有一些在未知样品中不存在的物种，而在未知样品中又含有一些在校准样品中不存在的物种的情况时，方程 (4.44) 和 (4.49) 都不成立 (即一个矩阵在另一矩阵的主成分上的投影将会改变前者的信息). 对 $[M]$ 和 $[N]$ 的加和进行主因子分析可以解决这一问题，定义 $[W]=[M]+[N]$，则

$$[W]=[\overline{U}]_W[\overline{S}]_W[\overline{V}]_W^T, \tag{4.53}$$

$$([\overline{U}]_W^T[M][\overline{V}]_W[\overline{S}]_W^{-1})[Z]_W^*=[Z]_W^*[\lambda]_W, \tag{4.54}$$

$$[Y]^T=([\overline{V}]_W[\overline{S}]_W^{-1}[Z]_W^*)^+, \tag{4.55}$$

$$[X][\beta]=[\overline{U}]_W[Z]_W^*. \tag{4.56}$$

特征值 $\lambda_k=\beta_k/(\xi_k+\beta_k)$. 对于同时存在于两种混合物体系中的所存成分，未知物中的浓度是 β_k，$\beta_k=\lambda_k\xi_k/(1-\lambda_k)$，当某一种物质不存在于校准样品中时，$\xi_k=0$，$\lambda_k=1$. 上述结论可用于前

面的所有情况，且不必用目标因子分析来进行检验. 为了计算，可通过向未知样品中加入含量已知的所有待分析物种而产生一个合成的 $[W]$ 矩阵，就是把校准混合样品加入到未知混合物中，直接量测得 $[W]$ 矩阵. 在实际应用中，可根据具体情况采取适当的获得 $[W]$ 矩阵的方法.

4.5 双线性目标因子分析

以上我们提到了双线性目标因子分析，下面我们对其作一简单说明. 同处理一维数据类似，目标因子分析也能处理双线性数据，对于检验向量 X_i 或 Y_i，目标因子分析可被表达为

$$[\overline{U}][\overline{U}]^T X_i = X_i$$

或

$$Y_i^T[\overline{V}][\overline{V}]^T = Y_i^T, \tag{4.57}$$

这里每一检验向量 X_i 或 Y_i 产生一预测的目标向量 X_i 或 Y_i. 如果检验向量存在于 $[M]$ 中 (即如果光谱为 $X_iY_i^T$ 的第 i 个组分存在于 $[M]$ 中)，那么预测的目标向量将等于该检验向量，因此

$$[\overline{U}][\overline{U}]^T X_i = X_i$$

或

$$Y_i^T[\overline{V}][\overline{V}]^T = Y_i^T, \tag{4.58}$$

应用 $[X]$ 和 $[Y]$ 的定义，类似地我们可写出

$$[\overline{U}][\overline{U}]^T[X] = [X]$$

或

$$[Y]^T[\overline{V}][\overline{V}]^T = [Y]^T, \tag{4.59}$$

此时，如果 $[N] = [X][\xi][Y]^T$，那么

$$\begin{aligned}&[\overline{U}][\overline{U}]^T[N][\overline{V}][\overline{V}]^T\\&= ([\overline{U}][\overline{U}]^T X)[\xi]([Y]^T[\overline{V}][\overline{V}]^T)\\&= [X][\xi][Y]^T = [N],\end{aligned} \tag{4.60}$$

也就是

$$[U][U]^T[N][\overline{V}][\overline{V}]^T = [N]. \tag{4.61}$$

这一方程定义了双线性目标因子分析. 对于

$$[\overline{U}][\overline{U}]^T[N] = [N]$$

和

$$[N][\overline{V}][\overline{V}]^T = [N], \tag{4.62}$$

由于随机噪音的存在，它们实际上都是近似的.

4.6 秩消因子分析法应用于萤光数据

C.N. Ho 等人首先将秩消因子分析法应用于多组分体系的激发 - 发射萤光光谱数据的定量分析，这些数据是由视频萤光计产生并直接收集于计算机软盘上的.

为了获得可适用于秩消因子分析的激发 - 发射矩阵 (EEM)，首先对萤光的 512 个视频帧进行加和，接着扣除相同数目的暗电流帧. 相类地以纯溶剂为空白，也采集一个散射光的 EEM，在进行数学分析之前，这个散射光 EEM 应从每一个 EEM 中予以扣除，激发单色器均处于低闪光状态，波长刻度分别设置在 381nm 和 462nm，激光和发射的入射狭逢均为 500μm, 研究所用试剂均经过区域纯化.

最先的工作是分析两套数据，一套是单组分芘的环己烷溶液，浓度变化范围为 1.0×10^{-6}M 至 1.0×10^{-9}M，另一套是芘和蒽的二组分环己烷溶液，其中芘的浓度变化范围为 1.0×10^{-6} M 至 5.0×10^{-9}M，蒽的浓度在全部 6 个样品中则维持在 2.0×10^{-5} M 附近. 对于单组分体系，根据浓度情况，设置不同的仪器增益进行数据采集；双组分体系数据的采集过程仪器增益维持不变，用于两个体系的标准是芘 5.0×10^{-8}M，蒽 1.0×10^{-6}M, 由萤光计所产生的 64×64 数据阵转入计算机后，缩小成一个 60×60 矩阵. 之所以这样处理，那是因为仪器本身在回描周期中所存在的内在局限使得原始数据矩阵中含有

少数几行和几列没有意义的数据，进行数学处理之前为方便起见，合并每一个 2×2 子块中的 4 个值成一个值，从而将数据阵缩小成一个 30×30 矩阵．扣除纯溶剂的空白样品的 EEM，以消除散射光的影响．

对各包含 6 个样品的上述两套数据进行秩消因子分析，所得结果同用最小二乘法分析所得的结果有好的一致性．然而，如果体系中有多种组分同时存在并且或者所存在的组分重迭严重，同时手头又缺乏所有的发射体的标准谱时，最小二乘法将很难提供出组分的正确浓度，此时，秩消因子分析却能在这方面显示出其优越性，甚至在对混合体系中的其余组分的本身都缺乏了解的情况下也可以对体系中的某一组分进行有效的测定．

为了更进一步证实秩消法的可应用性，C.N. Ho 等又将所测混合体系扩充至 6 组分，即芘、萤蒽、并四苯、二甲基蒽、蔻和蒽．这些多核芳香烃化合物的激发谱和发射谱显示出相当严重的重迭．共进行了 10 个混合样品的测定．混合样品采用以下方式来配制：①所有组分的浓度都成比例地变化；②当其他组分浓度维持不变时，改变蒽的浓度；③当维持其他 5 种组分相互之间的浓度比不变时，改变它们与蒽的浓度的比例．这次研究结果再次证明秩消因子分析对于定量多组分分析是一种强有力的工具，该法与以贝塞尔不等式为基础计算得的重迭度以及特征值对浓度作图等相结合，可大大增强人们对计算所得结果的确信度．值得引起注意的是，在应用秩消因子分析法去解决此类问题时，最严重的障碍是所研究的信号的清晰度和强度．

结合 Fletcher-Powell 的多元求最小算法，C.N. Ho 等对原提出的秩消因子分析法稍作改进以便能对混合体系中所有已知的组分进行同时计算．他们把改进了的这种秩消法称为同时多组分秩消法．应用该法对一套芘和并四苯的二元混合体系，一套蒽、并四苯和芘的三元混合体系和前面提到过的那套 6 组分多核芳香烃类混合体系进行同时计算，结果表明，对于同时存在的多组分的萤光分析来说，

该法是一种强有力的工具，具有计算速度快，效率高的特点，它使分析工作者具有选择样品中不同的组分数目乃至全部的组分数目来进行分析的灵活性，同传统的最小二乘法比较，这种方法也不受重迭谱的影响，它并不受一定要知道样品中的准确组成这样一种限制，而且，从对存在的物种作出不同的假设而得到的结果来看，方法本身具有良好的自我一致性.

4.7 秩消因子分析法在色谱中的应用

随着各种色谱检测技术的发展及计算机的广泛应用，快速采集各种经色谱分离后的组分的大量物理信息已成为可能. 液相色谱各馏分的传统光度测定及薄层色谱组分的紫外可见或萤光的光密度扫描，还有新近发展起来的光束成象分光光度计的应用以及光二极管阵列检测器的配备等都能最终获得具有双线性形式的数据矩阵. 因此，RAFA 技术在色谱法中便显示出极好的应用前景.

4.7.1 液相色谱

如果某化合物的色谱峰不与其他的色谱峰重迭，则这种化合物就易于定性和定量. 正因为如此，涉及色谱的分析工作大都着眼于如何获取分离得很好的色谱峰. 对于液相色谱来说，这往往要花费大量的时间和精力去确定影响一种或多种组分完全分离的色谱柱和流动相. 事实上，完全的分离是不易达到的，尤其是对于那些含有化学结构相似的组分的混合物来说，情况更是如此.

作为一种尝试，M. McCue 和 E.R. Malinowski 对液相色谱馏分的紫外吸光度进行秩消因子分析以达到单独地定量那些对未解析的色谱峰有贡献的组分的目的.

根据秩消因子分析法的原理，这种方法不要求对每一种组分都要有一个只有它自己才有吸收的波长. 虽然某一组分的色谱峰可能与混合物中的其他组分色谱峰产生重迭，但这种方法可帮助人们去

定量该组分而不需理会混合物中的其它组分的物种或含量.

为了易于理解,以紫外可见检测器的液相色谱为例来阐明RAFA在液相色谱法中对未解析峰的测定原理. 设标准溶液馏分的吸光值组成的数据矩阵为 $[N]$, 经色谱流动相稀释过的多组分未知溶液 (以下简称未知溶液) 的馏分的吸光值组成数据矩阵 $[M]$. 这两个矩阵的构成方式必须相同, 如行对应于不同的吸收波长, 而列对应于不同的馏分. $[N]$ 中的每一吸光值须符合比耳定律, $[M]$ 中的每一个吸光值必须是符合比耳定律的各个组分贡献的线性加和, 这样可得出下列表达式

$$M(i,\alpha) = b\sum_{j=1}^{n-1} a(i,j)y(j,\alpha)c(j) + ba(i,k)y(k,\alpha)c(k), \tag{4.63}$$

$$N(i,\alpha) = ba(i,k)y^0(k,\alpha)c^0(k), \tag{4.64}$$

式中, $M(i,\alpha)$ 是未知溶液的馏分 α 在波长 i 处的吸光值, b 为比色皿的光程长度, $a(i,j)$ 和 $a(i,k)$ 是组分 j 和 k 在波长 i 处的吸光系数, $y(j,\alpha)$ 和 $y(k,\alpha)$ 是未知溶液馏分 α 中组分 j 和 k 的分数, $c(j)$ 和 $c(k)$ 则是组分 j 和 k 在未知溶液中的浓度, $N(i,\alpha)$ 是标准溶液的馏分 α 在波长 i 处的吸光值, $y^0(k,\alpha)$ 则是组分 k 在标准溶液的馏分 α 中的分数, $c^0(k)$ 是组分 k 在标准溶液中的浓度, n 是未知溶液中的吸光组分数目. $\beta(k) = c(k)/c^0(k)$ 是一个要寻找的比例数. 因为 $c^0(k)$ 是标准溶液配制成时便已知的一个量, 如能找到 $\beta(k)$, 则未知的量 $c(k)$ 便可求得. 如果能适当控制色谱实验条件使得 $y(k,\alpha)$ 和 $y^0(k,\alpha)$ 相同, 那么, 就有可能找到 $\beta(k)$ 的值. 先重构一个修正的矩阵 $[H] = [M] - \beta[N]$, 这里, β 是一个标量, 将矩阵 $[H]$ 的秩看作是 β 的一个函数, 矩阵 $[M]$ 的秩应该等于未知溶液中的吸光组分数 n. 当 β 的取值恰好等于 $\beta(k)$ 时, 则有

$$M(i,\alpha) - \beta(k)N(i,\alpha) = b\sum_{j=1}^{n-1} a(i,j)y(j,\alpha)c(j), \tag{4.65}$$

这表明，组分 k 的吸光值贡献恰好已被从修正矩阵 $[H]$ 中消除，因此，此时的 $[H]$ 的秩也应较 $[M]$ 的秩小 1.

通过把修正矩阵 $[H]$ 的秩作为 β 的函数加以检验便可以研究这种秩变小过程. $\beta(k)$ 实质上就是能使 $[H]$ 的第 n 个特征值变为最小值时的 β 值. 很清楚，通过对修正矩阵 [H] 进行一系列的主因子分析就可以达到确定 $\beta(k)$ 的目的. 当然，为了确定最小的特征值，应该充分地变化 β 的取值.

通过将特征值对 β 值作图，可以更直观的看到第 n 个特征值的变化情形. 以 β 值为横坐标，以特征值为纵坐标，相对于每一个 β 值，标出共 r (是矩阵 $[H]$ 的行或列数，$r > n$) 个特征值，变换足够次数的 β 值后，连接所有第一个特征值成一条线，再连接所有第二个特征值成一条线，……，最后连接所有第 r 个特征值成一条线. 从图中就不难发现第 n 个特征值的连接线会出现最小值. 当然，由于实验误差的不可避免，故这个最小值是不会等于零的.

应用实例中，配制了 CH_3CN–H_2O 流动相 (351g CH_3CN/ 升), 标准乙苯 (1.38 ± 0.02 mg/ml), 标准邻二甲苯 (1.40 ± 0.02 mg/ml) 和标准对二甲苯 (1.03 ± 0.02 mg/ml) 溶液，同时还配制了两套含不同浓度的乙苯、邻二甲苯、对二甲苯的混合溶液作为未知溶液. 标准溶液和未知溶液的配制均以流动相为溶剂.

必须小心，为了保证 $y(k,\alpha)$ 和 $y^0(k,\alpha)$ 相同 (这是保证 RAFA 成功的关键), 要求应该用足够小的溶质浓度和应该有高度的色谱重现性. 足够小的溶质浓度可以保证在分离过程中被色谱分离的溶液中的该溶质具有独立的色谱行为，同时也可保证在每一组分的固定相浓度和流动相浓度之间存在线性关系. 为了获得与注入的溶质浓度无关的洗提体积和峰形，上述的色谱行为独立性和线性关系是需要的. 为了达到上述目的，在实验技术上可采用一个带有大容积样品旁路的注入阀以避免注入后柱头处瞬时形成最大的溶质浓度. 此外采用化学键合的固定相、精确地对色谱柱进行恒温、准确地测量洗提体积等可以保证足够的色谱重现性. 对于尖锐的谱峰来说，可

再现的保留数据特别重要，此时，洗提体积的一个小的变化将引起吸光值的较大的变化.

在这一应用实例中，在同样条件下分别收集校准溶液和未知溶液的 7 个馏分，用 VARIAN DMS 90 型紫外可见分光光度计来测量各馏分的吸光值. 测量时，带宽设置为 0.5nm, 用 2cm 比色皿，波长范围为 280—257nm, 测量间隔为 1nm (测得的吸收光谱如图 4.1 所示). 这样一来便可得到 24×7 校准溶液数据阵 $[N]$ 和未知溶液数据

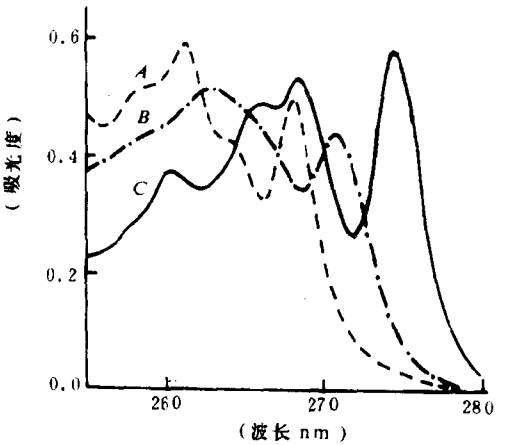


图 4.1 CH_3CN-H_2O 流动相溶剂 (每升溶液中含 351g CH_3CN) 的几种苯衍生物的吸光度光谱：(A) 乙苯；(B) 邻 - 二甲苯；(C) 对 - 二甲苯

阵 $[M]$. 对于每一个修正矩阵 $[H] = [M] - \beta[N]$, 都可算出相应的 7 个特征值. 将这些特征值作为 β 的函数来作图，就可得到类似图 4.2 的二维图. 从图 4.2 可以见到第三个最大的特征值作为 β 的函数显示出有最小值. 这是因为未知溶液中含有乙苯、邻二甲苯和对二甲苯，而校准溶液恰好含上述 3 种化合物中的一种 (乙苯) 的缘故. 对应于这一最小值的 β 就是我们要寻找的 $\beta(k)$ 值. 对两套含有不同浓度的乙苯、邻二甲苯、和对二甲苯的混合未知溶液的分析结果列于表 4.1 中.

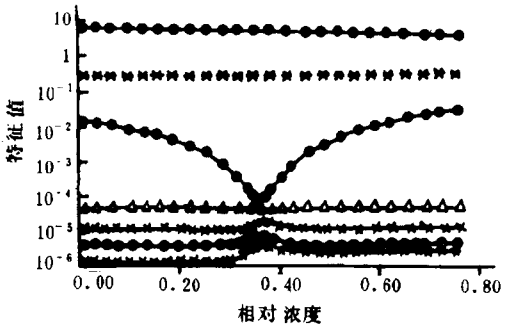


图 4.2 特征值作为相对浓度的函数 (未知溶液含有乙苯、邻二甲苯和对二甲苯，校准溶液只含有乙苯)

从表 4.1 中可见， RAFA 所得结果同配制时的加入量有很好的一致性. 由于在配制溶液时使用 2ml 的移液管，故在配制时就有可

表 4.1 秩消因子分析法结果与实验值比较

	在未知溶液 1 中的重量分数		在未知溶液 2 中的重量分数	
溶质	加入的	RAFA 结果	加入的	RAFA 结果
乙苯	0.249 ± 0.004	0.239	0.599 ± 0.007	0.594
邻二甲苯	0.505 ± 0.006	0.498	0.203 ± 0.003	0.202
对二甲苯	0.247 ± 0.004	0.246	0.198 ± 0.003	0.201

能产生一定的误差，所以，上述结果表明方法所设计的色谱技术和操作步骤完全符合 RAFA 技术的要求.

4.7.2 薄 层 色 谱

为了提高获取大量信息的速度和提高工作效率， M.L. Gianelli 等人装配一种称为单光束成象分光光度计的仪器用以在薄层色谱板上对混合物进行直接分析，这种仪器可以在数分钟内记录目标板上的多达 64000 个位置上的反射谱或透射谱，对薄层色谱板上各个点用不同波长的单色光进行透射，再通过光导摄象管摄象机和计算机连接进行实时数据采集，然后，按一定的移动方向集合数据，最终便能获得具有双线性形式的数据矩阵.

如果想分析未知混合物中的某些组分而对其余组分又完全无所了解时，秩消因子分析技术是比较合适的．他们应用 RAFA 去分析四苯卟吩 (简称 H_2TPP), 四苯卟吩锌 (ZnTPP) 和初卟吩钯 (PdEP) 的混合物．为了在薄层板上成斑，它们都被溶解于二氯甲烷中，它们的色谱斑点严重重叠，表 4.2 中列出了它们分别用秩消因子分析法 (RAFA) 和最小二乘法 (LS) 分析二维薄层色谱数据的结果．

表 4.2　RAFA 法和 LS 法分析二维薄层色谱数据比较

溶质	加入量 (μg)	RAFA 法结果 (μg)	LS 法结果 (μg)
H_2TPP	1.40	1.57	1.69
ZnTPP	0.91	0.86	0.73
PdEP	1.37	1.44	1.46

从表 4.2 中可以看出，秩消因子分析能给出更满意的结果，由此可见，秩消因子分析法对于薄层色谱斑点重迭严重的情况是一种值得优先考虑的方法．尽管其它的色谱法较传统的单向薄层色谱具有更高的分辨率，但是后者固有的大的适应性和轻便性以及二维薄层色谱法的易于完成，使得人们更加喜欢采用薄层色谱法和吸光度检测．特别是，分析所需费用较低，又能平行操作等这些优点更值得分析工作者考虑，秩消因子分析法等数学技术的应用，更加拓宽了薄层色谱法的应用范围．

4.8　秩消因子分析法在其它方面的应用

B.E. Wilson 等人报导了他们用计算机模拟将秩消因子分析技术用于二级非线性数据的研究工作，他们还将工作延伸到对 D_2O 中 6 个糖的 2DJ- 偶合 NMR 谱的研究．在有干扰谱存在的情况下，他们将该法同其它 3 种曲线解析方法就其准确地预测被测物的能力进行比较 (多元线性回归被用作参照方法), 结果认为在试验的几种方法中，非线性秩消法是唯一具有解决实际化学问题潜力的方法．

5 渐进因子分析

5.1 概　　述

通过前面章节的介绍，我们已经知道，抽象因子分析 (AFA) 揭示了在一系列含有相同组分但各组分的比例不同的有关混合物中所存在的组分的数目. 目标因子分析 (TFA) 可确认被怀疑为存在于混合体系中的组分是否存在. 秩消因子分析 (RAFA) 在不借助对其他组分有任何了解的情况下，可对特定组分进行定量. 然而，对所含组分不明，组分含量及其组分数未知的混合物的化学分析，却是分析工作者所面临的最困难的问题之一. 根据光谱滴定的平衡研究和色谱中的峰分辨是通过采用多波长 (更一般地说是多通道) 检测系统可以合适地得到解决的两类典型问题. 例如，在 w 个波长处测量 m 个谱构成的一个 $m \times w$ 数据矩阵 $[Y]$. 在这些研究中，处理数据的最主要目的就是把原始数据矩阵 $[Y]$ 分解为两个较小的矩阵，即由各组分的浓度构成的矩阵 $[C]$ 和吸光系数矩阵 $[A]$. 然而，在数学上，分解 $[Y]$ 不是唯一的. 在平衡研究中，分析一般是以某种化学模型 (即关于化学型体的种数、种类的假设和质量作用定律) 为基础的. 但在许多情况下，挑选正确的、合理的化学模型决不是轻而易举的事情. 况且质量作用定律是不能应用于诸如色谱中的峰分辨或是以具有强烈变化的离子强度或溶剂成分的测量为基础的平衡体系的分析等这一类问题的. 因此，一种不那麼依赖化学模型而又能阐明各单个组分的分布轮廓的方法理所当然会受到注意.

近年来，出现了一类被称为渐进因子分析 (Evolutionary Factor Analysis，简称 EFA) 的方法，它包含那些利用在化学中发生的渐进过程的因子分析技术. EFA 大体上可分成两种：一种是有模型的，

另一种是自模或无模型的. 本章中我们只介绍和讨论第二种. EFA 能充分利用原始数据直接包含的信息，具有独特的优点，它既不要求所怀疑的组分的光谱，也不要求每一种组分有独特的波长点，而是利用每一种型体在它的渐进浓度分布曲线中有一单独、唯一的最大这一事实. 与某些其他方法比较起来，这类方法不限于两组分或 3 组分体系. 这些优点将在我们叙述它的应用时充分地得到体现. EFA 总的思想是采用渐进方式处理数据矩阵，但在具体的处理过程中采用的手段是不尽相同的. 据此，大致上把它们归纳为 3 种方法 (用提出者的姓名命名)，即 Gemperline 法，VDK 法和 GMMZ 法. 下面，我们简单介绍一下这 3 种方法的基本思路.

5.1.1 Gemperline 法

P.J. Gemperline 提出了一种自模 EFA 技术，用于解释高效液相色谱 (HPLC) 的重叠峰. 该法既不要求有可疑组分的纯光谱，也不要求对每一组分显示出唯一响应的波长点.

为搞清楚这种方法的原理，让我们来考虑一个由紫外 / 可见吸光度组成的 $r \times c$ 数据矩阵，这些吸光度是在 r 个波长点和跨越一个未被分辨的色谱带的 c 个数字化了的时间间隔下测得的. 采用一种配备光二极管阵列检测器 (DAD) 的商用高效液相色谱可得到这类数据矩阵. 利用 AFA 可自动地将这一矩阵分解成一个 $r \times n$ 抽象的吸收系数矩阵 $[E]_{\text{abst}}$ 和一个 $n \times c$ 的抽象浓度矩阵 $[C]_{\text{abst}}$

$$[A] = [E]_{\text{abst}}[C]_{\text{abst}}, \tag{5.1}$$

n 是吸光组分数，n 的确定有多种方法，前面已作过介绍. 方程式 (5.1) 假设所有条件均满足比耳定律的要求.

$[E]_{\text{abst}}$ 和 $[C]_{\text{abst}}$ 是符合式 (5.1) 的纯数学解. 然而，遗憾的是，它们缺乏化学意义，所以，有必要将这些矩阵变换为所研究的组分的有物理意义的吸收系数和浓度矩阵. 从第三章中的介绍知道，这可由目标因子分析 (TFA) 来完成. TFA 是一种涉及到对模拟组分的光谱或洗提曲线向量进行的目标检验技术. TFA 的效能在于其

检验是用每一个可疑的目标向量逐个地进行的，不需要预先了解与其他组分有关的信息.

目标检验把洗提曲线检验向量 C_{test} 转变为一个完全处于因子空间中的预测向量 C_{pred}. 通过下面的计算使 C_{test} 和 C_{pred} 之间差的平方和达到最小就可做到这一点，即

$$C_{\text{pred}} = C_{\text{test}}[C]^T_{\text{abst}}\{[C]_{\text{abst}}[C]^T_{\text{abst}}\}^{-1}[C]_{\text{abst}}, \tag{5.2}$$

当检验向量和预测向量较好地吻合时，则正确的检验向量被鉴别出来. 当 n 个这样的向量被找到后，就把它们当作向量而组合成真实浓度矩阵 $[C]_{\text{real}}$，真实的吸收系数矩阵 $[E]_{\text{real}}$ 可通过下面的关系式求得

$$[A]_{\text{real}} = [E]_{\text{real}}[C]_{\text{real}}, \tag{5.3}$$

即

$$[E]_{\text{real}} = [A]_{\text{real}}[C]^T_{\text{real}}\{[C]_{\text{real}}[C]^T_{\text{real}}\}^{-1}. \tag{5.4}$$

$[E]_{\text{real}}$ 的每一列描绘出一个真实组分的光谱.

通过将一个元素的值设为 1，其余各元素值置零的办法可构造出独特性检验向量，一个完整的独特向量集，对每个数字化的洗提时间 t_j 包含有单位 1，即

$$\left.\begin{aligned} C_{\text{t1,test}} &= (1,0,0,\cdots,0,0),\\ C_{\text{t2,test}} &= (0,1,0,\cdots,0,0),\\ &\cdots\cdots\cdots\cdots,\\ C_{\text{tc,test}} &= (0,0,0,\cdots,0,1). \end{aligned}\right\} \tag{5.5}$$

这些检验向量代表了虚设的洗提曲线. 按照 Gemperline 法，如果以 $C_{tj,\text{test}}$ 表示的保留时间与某一真实组分的保留时间相对应，则 $C_{tj,\text{test}}$ 和 $C_{tj,\text{pred}}$ 之间的差值将会达到一局部极小. 因为预测曲线是一个比虚构的“独特”曲线更好的描述，如果发现多于 n 个的局部极小值，则仅选择那些含有 n 个最小的极小值的向量.

负的区域在物理上是无意义的. 通过舍去所有超越峰值最大的左边和右边所遇到的第一个负区域为标志的界限的数据便可精化这

些曲线. 用迭代方法可产生新的曲线，迭代进行至没有太大意义的精化出现为止. 在出现下面两种情况之一时可终止迭代：①当目标中的真实误差 (RET) 小于预测向量中的真实误差 (REP)(误差函数来自目标因子分析中的误差理论) 或②当在连续的迭代之间，被舍去的点的误差下降变化比预测误差 (REP) 小时.

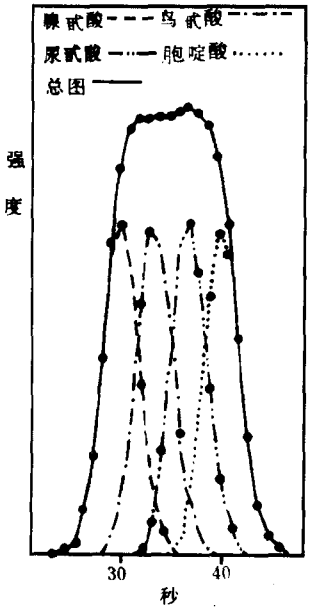


图 5.1 Gemperline 法预测的混合酸组分的洗提曲线

Gemperline 把上述的解析技术试用于腺甙酸，胞啶酸，鸟甙酸和尿甙酸的各种模拟混合物. 模拟数据矩阵由高斯洗提曲线和紫外光谱构成，为了研究实验误差的影响，随机数被加到数据矩阵中，解析结果与用以生成数据矩阵的原始曲线有很好的吻合，如图 5.1 所示.

5.1.2 VDK 法

B. Vandeginste， W. Derks 和 G. Kateman 报导了一个与 Gemperline 法类似的方法. 该方法采用了多核芳烃混合物和蛋白质混合物的 HPLC-DAD. 这些研究者对抽象浓度矩阵进行方差最大旋转，这是一种正交旋转，它使载荷的平方的总方差最大，使用这一方法沿着未知浓度曲线尽可能好地调整抽象因子，鉴别出组分的保留时间 (峰最大). 这样，对应于这 n 个已被鉴别出的保留时间的独特性检验向量便可建立并通过迭代而使其精化. 为保证收敛于纯曲线，通过置零来精化每一条曲线：①出现小于某给定阈值 (如 0.005) 的任何元素被置零；②由一个或多个零分隔的任何双峰中的较小峰值

被置为零. 在出现下列情况中的任何一个时，停止迭代：①目标不能被进一步精化；②检验向量与预测向量之间的相关系数超过预先给定的值；③迭代的次数达到指定的最大值.

除了用以鉴别开始的独特性单值向量的初始手续不同外，此法在本质上与 Gemperline 法相同.

5.1.3 GMMZ 法

另一种 EFA 方法是由 H. Gampp， M. Maeder， C.J. Meyer 和 A.D. Zuberbuhler 等提出的. 这一无模型方法是以对在渐进过程中得到的一组数据矩阵重复进行特征分析为基础的. 对完整的数据矩阵序列 (这些矩阵是在渐进过程中通过连续地将光谱加到前面的矩阵中而构成的) 进行特征分析，当新的吸光物种出现时，抽象因子的特征值增加一个数量级，这一过程称为正向渐进因子分析.

反向渐进因子分析是这样进行的：从最后的两行光谱数据来开始因子分析，然后按收集数据的相反顺序连续地将光谱数据逐个地加到矩阵中进行特征分析. 结果所得到的一个较大的特征值可检测出一个组分的消失.

由正向分析和反向分析得到的特征值作为渐进变量 (如 pH) 的函数在同一图上绘图，在第 i 个正向特征值曲线和第 $(n+1-i)$ 个反向特征值曲线下面的两条曲线共有区域描述出第 i 种型体的浓度曲线. 图 5.2 的 (b) 绘出根据分光光度研究 Cu^{2+} 与 3.7-diazanonane 反应的 pH 滴定数据所得的典型结果. 由于体系中存在 4 种组分，n 等于 4，因此，为了得到浓度曲线，连接特征曲线 (正向 – 反向) 如下： 1–4′, 2–3′, 3–2′ 和 4–1′, 结果所得到的浓度曲线同直接从电子自旋 (ESR) 数据所得到的曲线 (图 5.2 的 (a)) 非常一致.

为了更进一步精化曲线， H. Gampp 等提出如下迭代处理：首先，归一化浓度曲线，然后用方程 (5.4) 由归一化的浓度曲线计算光谱曲线. 设所有负的吸光度值为零，应用已校正过的吸光度按下面的方程重新计算浓度矩阵 (即浓度曲线)

$$[C_{\text{real}}] = \{[E]^T_{\text{real}}[E]_{\text{real}}\}^{-1}[E]_{\text{real}}[A], \tag{5.6}$$

再将计算结果中所有的负的浓度值置为零，规一化每一条曲线，然后重算光谱矩阵 $[E]_{\text{real}}$，重复上述过程直至收敛为止.

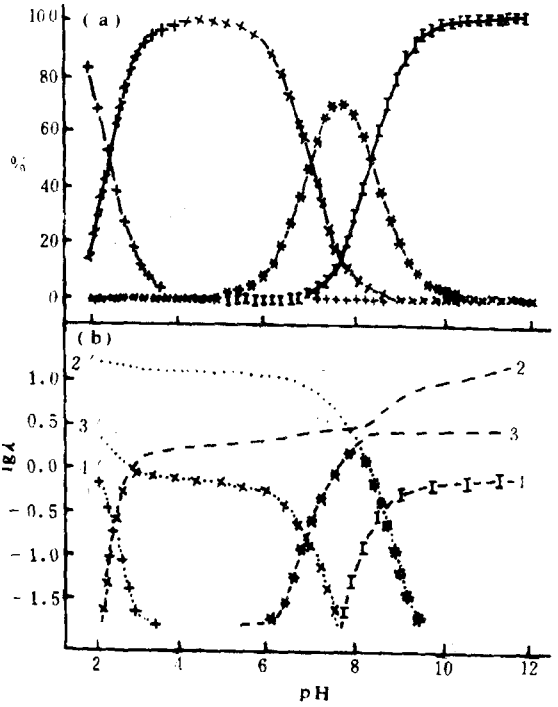


图 5.2 Cu^{2+}-3.7-diazanonan-diamine 作为 pH 的函数 (a) 电子自旋共振数据给出的物种分布曲线；(b) 可见光谱数据的正向和反向 EFA 产生的特征值图

5.1.4 几种方法的比较

虽然上面介绍的 3 种方法看起来似乎并不相同，但它们却有着许多共同之处. 它们的基本差异在于用来产生潜在组分的原始浓度

曲线的原理不同，在每一种情况下，原始的未经精化的曲线都能正确地鉴别该组分的浓度最大值．然而，每一种方法在初始阶段得到了相当不同的曲线，这些曲线需要基本的迭代精化处理．

用 Gemperline 或 VDK 法所得到的初始曲线往往显示出不希望有的特征，如①异常大的负浓度；②双最大值，最大值之间没有零基线区域；③在超越基线边界的区域中的不合理的峰等．曲线的最终形状对迭代精化过程所采用的判据是敏感的．

GMMZ 法比较费时，它要求正向和反向特征值曲线的正确配合，这并不是一个无足轻重的问题．如果处理不当，最终的结果将是无意义的．然而，该法能以合理的精度确定对应于组分的出现或消失的洗脱位置，这一信息在其他两种方法中都是没有的．因此，虽然从 GMMZ 法所获得的初始曲线是比较粗糙的，但用 Gemperline 判据去进行精化应该说是最成功的．这一点将在圆二向谱的有关应用中得到证实．因为 EFA 在分析完全未知的络合物、混合物的问题时能给出合理的答案，所以，EFA 应该被认为是一类有发展前景的技术，它们对组分的光谱相似性和相异性以及组分的洗提可分离性是敏感的．不过，这些方法仍须通过努力使其能够确定组分的浓度曲线和组分的分离光谱置信度和误差界限．

5.2 渐进因子分析的原理

上一节，我们已经简单介绍了几种渐进过程方法．鉴于 GMMZ 法系统性较强，本节将以处理光谱滴定数据为例，介绍 H. Gampp 等人提出的渐进因子分析法的数学概念．

5.2.1 方法的数学描述

因为这种无模型的渐进处理方法是以主因子分析为基础的，而且把独立主组分 (即因子) 的逐渐形成作为滴定过程的函数，故称之为渐进因子分析．

在分析光谱数据时，不仅要确定体系中存在的吸光物种，而且还需要确定各吸光物种的吸收光谱及其相应的含量 (浓度). 光谱滴定分析的首要任务是把测得的原始吸光度矩阵 $[Y](m \times w)$ 分解为两个较小的矩阵，即浓度矩阵 $[C](m \times s)$ 和吸光系数矩阵 $[A](s \times w)$，其中 w 为测定的波长点 (通道) 数， s 为吸光物种数， m 为谱图个数 (如对应于 m 个不同的 pH 值可得 m 个谱图). 解决这一问题的经典方法是通过对给定的化学模型寻找最佳的非线性最小二乘拟合来实现的，最小二乘法需要选择合适的化学模型，然而，这恰好是一件不容易的事情.

对于一给定包含 s 独立吸光物种信息的 $m \times w$ 光谱数据矩阵 $[Y]$ 可用主成分分析法将其分解为抽象的 $m \times s$ 行矩阵和 $s \times w$ 抽象列阵的乘积

$$[Y] = [L][E], \tag{5.7}$$

也可通过对 $[Y]$ 进行奇异值分解得到 3 个矩阵变换形式

$$[Y] = [U][S][V], \tag{5.8}$$

式 (5.7) 和 (5.8) 有着紧密的联系：其中 $[L] = [U][S]$， $[V] = [E]$，$[S]$ 是对角矩阵，其对角元素是 $[Y]^T[Y]$ 的按降序排列的特征值矩阵 $[\wedge]$ 中元素 λ_i 的正的平方根， $[\wedge]$ 是 $[Y]^T[Y]$ 的 S 个较大的主要特征值为对角元素组成的矩阵. $[U]$ 或 $[L]$ 各自的列之间以及 $[V]$ 或 $[E]$ 各自行之间是正交的. 即

$$[U]^T[U] = [V][V]^T = [E][E]^T = [I], \tag{5.9}$$

$[I]$ 是单位矩阵. 以及

$$[L]^T[L] = [\wedge] = [S^2], \tag{5.10}$$

$[U]$ 是由 $[Y][Y]^T$ 的主要的 (s 个) 特征值对应的特征向量组成. $[V]$ 是 $[Y]^T[Y]$ 的主要的 (s 个) 特征值对应的特征向量组成.

EFA 最终目的是将 $[U]$ 和 $[V]$ 组成的抽象的向量空间转化为由真实浓度 $[C](m \times s)$ 和吸收光谱 $[A](s \times w)$ 构成的向量空间，即

$$[U][S][V] = [C][A] \tag{5.11}$$

或

$$[L][E] = [C][A],$$

相对于一般的因子分析来说，应用 EFA 有一个附加的重要的条件，即原始数据矩阵必须以这样的方式来排列，即浓度分布是相切的．换句话说，原始数据 $[Y]$ 是由 m 个量测光谱以下述方式来构成的：每一物种仅与 m_i 个量测有关 ($m_l \leq m_i \leq m_u$，m_l 和 m_u 分别表示量测的低限和高限) 且在有关的量测低限和高限范围之外时，便失去考虑的价值．见图 5.3.

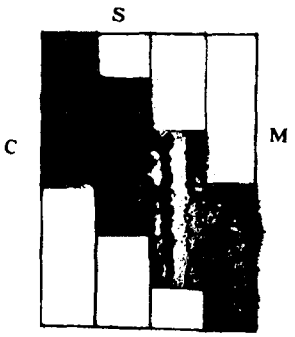


图 5.3　在一个包含有 4 个物种的模型中有关浓度的相切范围

EFA 主要包括两个独立的部分，第一部分称为初级渐进因子分析，第二部分称为终级渐进因子分析．

5.2.2　初级渐进因子分析

初级渐进因子分析的基本思想是通过逐个地对 $[Y]$ 的子阵进行主因子分析或奇异值分解来观察 $[Y]$ 秩变化及其伴随产生的 $[Y]^T[Y]$ 的特征值与滴定过程变化的关系．它包括下面几个步骤：

(1) 首先对原始数据矩阵 $[Y]$ 进行主因子分析或奇异值分解便可产生 s 个有意义的因子，$[L]$ 和 $[E]$，以及残余偏差 δ_y．δ_y 可通过下式求得，即

$$[R] = [Y] - [L][E],$$

$$\delta_y = (\sum\sum R^2(i,j)/(m \times w - s))^{1/2}. \tag{5.12}$$

(2) 对 $[Y]$ 进行主因子分析之后，逐次地由 $[Y]$ 计算二次矩阵

$$[M_i] = [L_i]^T[L_i], \quad (i = 2, 3, \cdots, m). \tag{5.13}$$

然后对 $[M_i]$ 进行主因子分析，求其特征值构成的 $[\wedge_i]$. 其中 $[L_i]$ 是由 $[L]$ 的前 i 个行向量构成的 $i \times s$ 矩阵，$[M_i]$ 也可通过 $[M_i] = [Y_i]^T[Y_i]$ $(i = 2,3,\cdots,m)$ 求出. 值得注意的是，由 $[L_i]^T[L_i]$ 组成的 $s \times s$ 矩阵 $[M_i]$ 的特征值和由原始数据组成的 $w \times w$ 矩阵 $[Y_i]^T[Y_i]$ 的 s 个较大的特征值是相同的. 然而，用由 $[L_i]$ 计算所得的 $[M_i]$ 求特征值的工作量比用 $[Y_i]^T[Y_i]$ 求特征值的工作量要小得多. 另外，无论是用哪种方式构成 $[M_i]$ 及计算特征值 $[\wedge_i]$，使用下面的方法都能大大地提高效率. 即从前一步 $i-1$ 开始，$[M_i]$ 的每个元素 $M(j,k)$ 可简单地由下式得到

$$M_i(j,k) = M_{i-1}(j,k) + L(i,j)L(i,k) \tag{5.14}$$

或

$$M_i(j,k) = M_{i-1}(j,k) + Y(i,j)Y(i,k),$$

$L(i,j)$ 和 $L(i,k)$ 分别是 $[L]$ 的第 i 行的第 j 个和第 k 个元素，$Y(i,j)$ 和 $Y(i,k)$ 的含义可照此类推. 通常，由于前一步骤的第 $i-1$ 个特征向量为采用向量迭代精化特征值 $[\wedge_i]$ 提供了很好的初始值，所以 $[\wedge_i]$ 的计算也相应地变得很快. 初级 EFA 的这一部分称为正向 EFA. 将分析得到的诸特征值 $[\wedge_i]$ 组合成一个 $m \times s$ 渐进因子或浓度矩阵 $[C_f]$. 因为 $[Y_i]$ 是 $[Y]$ 的子阵 ($[L_i]$ 也类似)，它包括 $[Y]$ 的前 i 个光谱 (或称行)，这样，当新的物种开始形成并变得有意义 (即对数据有贡献) 时，一个新的较大的特征值 (因子) 就会逐渐出现.

(3) 把计算得到的特征值 $[\wedge_i]$ 作为滴定过程的函数作图. 因为特征值通常跨越几个数量级，所以取其对数值更适合于图形的表示，如图 5.4(a) 中的 - - - 线所示. 一个较大的有意义的特征值出现就意味着一个新的吸光物种在滴定过程中形成，正向 EFA 图形不仅显示了子矩阵 $[Y_i]$ 包含不同物种数目的信息，而且，由此可看出特征值的大小在一定程度上和物种的浓度有关. 当然，一个给定的特征值 (甚至是它的对数值) 并不会直接对应于一个给定的络合物或其浓度，实际的特征值在很大程度上取决于各种吸光物质的非相似性以及它们的浓度之间的差别. 尽管如此，在浓度和特征值之间毕竟存在着

$[R]=[S]([A][V]^T)^{-1}$ 是式 (5.11) 的一个非正交旋转或者说是 $[C]$ 在 $[U]$ 上的投影. 式 (5.18) 说明浓度曲线 ($[C]$ 的列向量) 是浓度特征向量 ([U] 的列) 的线性组合. 对于一特定的组分浓度 C_i(向量)，可写成

$$C_i=[U]R_i, \tag{5.19}$$

这里的 C_i 是 $[C]$ 的第 i 列向量，R_i 是 $[R]$ 的第 i 列向量. R_i 的计算是以由初级 EFA 产生的浓度窗口为基础的. 等式 (5.18) 和 (5.19) 包含的特征示于图 5.5 中. (a) 部分的 $[C]$ 中对应的列的无阴影部分表示每个组分的范围，在这些窗口以外的阴影部分，$[C]$ 的元素被记为零. 图 5.5 中 (b) 代表等式 (5.19). 这里的关键思想是把 C_i 的阴影部分看成是 $[U]$ 中阴影部分的线性组合，现在我们把着眼点放在阴影部分便可得到图 5.5 的 (c) 部分. C_i^0 是一个零向量. 因此这是一个具有明显无效解 $R_i=0$ 的齐次方程组. 当组分 C_i 消失时，$[U^0]$(含 N 行) 的秩仅为 $N-1$. 因此，这个齐次方程组具有非无效解：R_i 的一个元素 $R_{i,1}$ 可随意选择，为方便起见，可固定为 1. R_i 的其他元素，如 $R_{i,2}$ 可通过下面的线性回归计算出来，即

$$\begin{aligned}R_{i,2}&=-([U_2^0]^T[U_2^0])^{-1}[U_2^0]^T[U_1^0]R_{i,1}\\&=-([U_2^0]^T[U_2^0])^{-1}[U_2^0]^T[U_1^0],\end{aligned} \tag{5.20}$$

式中的 $[U_1^0]$ 是对应于 $R_{i,1}$ 的 $[U^0]$ 的子集，$[U_2^0]$ 是对应于 $R_{i,2}$ 的 $[U^0]$ 的子集.

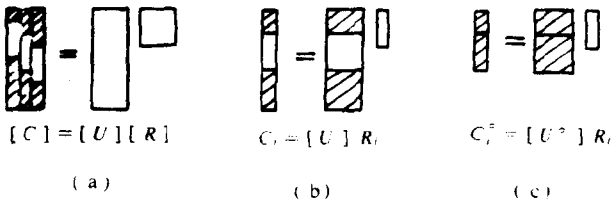


图 5.5 式 5.18 和 5.19 说明示意图

在求得 R_i 之后，由方程 (5.19) 求出 C_i，重复上述过程直至求得所有物种 (s 种) 的浓度为止，最后由诸 C_i 构成矩阵 $[C]$，再由 $[C]$ 按下式求出吸收光谱矩阵 $[A]$. 即

$$[A] = ([C]^T[C])^{-1}[C]^T[Y]. \tag{5.21}$$

由上述步骤可以看出，非迭代的 EFA 处理要比迭代的 EFA 的处理快，就方法而言，迭代的 EFA 是通过反复使用目标因子分析而获得 $[C]$ 和 $[A]$ 的；非迭代的 EFA 是在浓度窗口的基础上直接使用最小二乘方法. 两者各有特色，后者在用于色谱峰的分辨中获得满意的结果. 在实际应用中可根据具体情况选用合适的方法.

5.3 其他无模型技术及其与渐进因子分析的比较

大多数用于光谱和浓度分布曲线的无模型估算方法都以 W.H. Lawton 和 E.A. Sylvestre 的开拓性工作为基础. 他们的自模曲线分辨仅仅以吸收光谱的非负性这一事实为基础，其主要缺点是只局限于两组分体系. 自模曲线分辨技术已成功地用于色谱和 X 射线光电子谱等研究领域.

其后，又有将该法扩展到 3 组分或更多组分的报导. 在这些情况下，有些困难需要加以克服，如由特征向量集 $[U]$ 或 $[V]$ 所旋转的多维空间的图形显示就是棘手的. 一种可能性是转化为极作标 (这会使浓度方面的信息变得模糊)，或者映射到超平面上. 另外，只要问题解的可接受范围仅为物理参数 (浓度和吸光系数) 的非负性所定义，那么，对于这样的多组分体系，所得结果的意义通常是不够明确的，有几种进一步的约束，诸如所有光谱或浓度分布图最大相异性或最小包线被提出用来获得问题的独特解. 但是，关于这样的假设一般来说是没有物理事实根据的.

最近，已经发展了几种用于对吸收光谱和浓度分布的初始估计值进行迭代精化处理的无模型解析的方法. 这些方法对组分的最大数目没有内在的限制，绝大多数的这些算法是以迭代的目标因子分析 (ITFA) 为基础的. ITFA 的解题策略可分为两部分，首先必须选出浓度分布曲线的一组近似值，第二步通过重复目标变换用迭代方式来精化这些近似值

$$[C_{\text{new}}] = [U][U]^T[C_{\text{old}}]. \tag{5.22}$$

已设计不同的方法来调整得到的浓度分布 $[C_{\text{new}}]$(这是保持迭代进行所必须的)，所有这些方法在每次迭代循环中都把 $[C]$ 中的负值置零，在 P.J. Gemperline 的有关工作中，这是唯一的校正. 如图 5.6 所示，对用来检验 EFA 的模型数据，采用迭代目标变换所得的结果 (虚线表示) 偏离原始的真实浓度曲线. 有一种方法可以防止这种偏离，该法把浓度矩阵 $[C]$ 中的“小”的正值也置为零，这个方法已成功地用于含多至 6 个重叠组分的色谱峰的解析，显然，选择合适

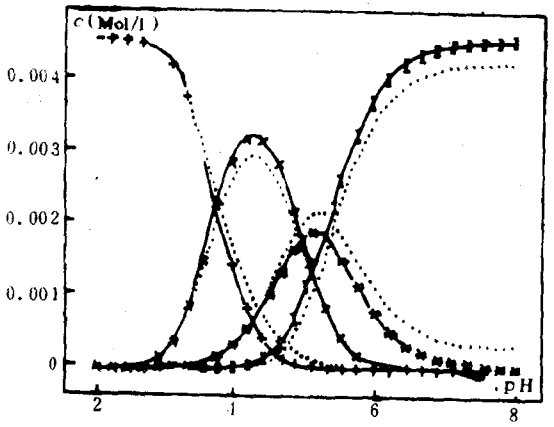


图 5.6 模型数据的浓度曲线

—— 为迭代 EFA 结果 ····· 为迭代目标变换结果 (+ : Cu^{2+}, × : CuL^{2+}, ⋆ : $CuLH^{+}_{-1}$, I : $CuLH_{-2}$)

“小”的程度是重要的，可惜的是还没有通用程序．使用正值的截止水准当然可以帮助避免不恰当的偏离，然而，这一水准的选取似乎要经过反复的试验，并且对于是扩大还是缩小，这种算法仍无特有的防错措施． R.F.Lacey 提出类似的方法，该法把统计误差较大的浓度值重新置零．这个问题可通过使用 EFA 得到解决，体系中所有物种的存在范围都可被唯一地确定下来．综上所述，我们觉得 EFA 与其他新的无模型方法的主要区别不在于迭代精化的细节上，而是在于是否利用通过初级 EFA 而得到的“浓度窗口”这一信息．

5.4 渐进因子分析的应用

5.4.1 浓度、光谱及平衡常数的计算

同其他平衡研究一样，光谱滴定数据处理通常包括以化学模型为基础的确定平衡常数的最小二乘精化，使用相应的平衡常数的选定模型估算值和组分的分析浓度可计算浓度分布曲线，这些曲线是不能以任何方式直接从实验数据获得的．摩尔消光系数甚至可从算法中删去而在精化的最后阶段不经迭代便算得．近 10 多年来，在这方面已取得了重大进步．功能强的程序可处理光谱数据，这几乎象处理电位滴定数据那样寻常，甚至对于更复杂的体系也是如此．我们不打算在这里讨论电位滴定和光谱滴定各自的优点，但是，目前的技术发展水平似乎表明光谱滴定可与电位滴定在可靠性及精度方面相媲美，而在区别不同的化学模型的能力方面，一般地说，光谱滴定更具优势．

渐进因子分析在计算浓度分布曲线、吸收光谱和平衡常数等方面开辟了一条完全不同于经典方法的新途径．其初始步骤是把测量的吸光度数据矩阵 $[Y]$ 分解为浓度分布矩阵 $[C]$ 和摩尔消光系数矩阵 $[A]$，对 $[C]$ 和 $[A]$ 直接进行最小二乘精化，使残余偏差 $[R]$ 的平方最小化，即

$$[R] = [Y] - [C][A], \qquad SQ \overset{\min!}{=\!=} \sum_{i=1}^{m}\sum_{j=1}^{w} R(i,j). \tag{5.23}$$

在完成这一数据处理之后，再借助相应的吸收光谱和 (或) 有关化学理论确定与浓度分布有关的化学物种，平衡常数在最后算得，即在给前面的抽象物种确定化学计量数后直接从浓度分布曲线的各自截距读出. 应用 EFA 时，处理光谱滴定问题的正常次序就被改变了，首先按 $[Y] = [C][A]$ 分解数据，然后选择一个模型，最后引入质量作用定律.

EFA 的这种应用业已经过分光光度、 ESR 和半合成的模型数据的广泛验证，一般说来，结果是满意的，甚至对于 4 组分或 5 组分体系，情况也是如此，在大多数情况下， EFA 和质量作用定律所算出的浓度之间的标准偏差在 0.5—2% 之间，所计算得的吸收光谱实质上是一致的，最后所算出的稳定常数也在实验误差范围内相吻合. 用 EFA 计算出的数据与实验数据之间的误差甚至比用质量作用定律时要小，这是令人满意的. 对于这种成功的分析不必感到意外，因为 EFA 没有更多的严格的限制，因而就更容易适应有实验误差的分析.

EFA 的功能可用一个典型的例子予以证明，如图 5.7 所示. 数据产生于由 Cu^{2+} 及其络合物 CuL^{2+} ， $CuLH^{+}_{-1}$ 和 $CuLH_{-2}$ (L 为 $H_2NCOCH_2NH(CH_2)_3NHCH_2CONH_2$) 这 4 种吸收严重重叠的物种所组成的混合物体系. 其中，与实验测量对应的随机误差 (2×10^{-4} 吸光单位和 0.008pH 单位) 被叠加在数据中. 图中显示了 3 种浓度分布曲线：①初级 EFA 的结果 (虚线) ；②精化处理后的 EFA 结果 (实线) ；③基于质量作用定律的分析 (符号线). 由此图可见，后两个结果实际上是相同的. 从 EFA 图中可读出正确的平衡常数：$\lg K^{Cu}_{CuL} = 9.99(9.99)$ ， $\lg K^{H}_{CuL} = 5.92(5.33)$ ， $\lg K^{H}_{CuLH_{-1}} = 5.02(5.01)$ (括号中的结果是由质量作用定律获得的). 这种结果并不奇怪，因为 EFA 的结果是通过对在 pH 为 3.8 ， 5.3 和 5.2 附近的相应的交叉

点进行线性插值而得到的，对于游离的配位体的质子常数，采用的是 $\lg K^{H}_{LH} = 8.40$， $\lg K^{H}_{LH_2} = 6.55$ 等数据.

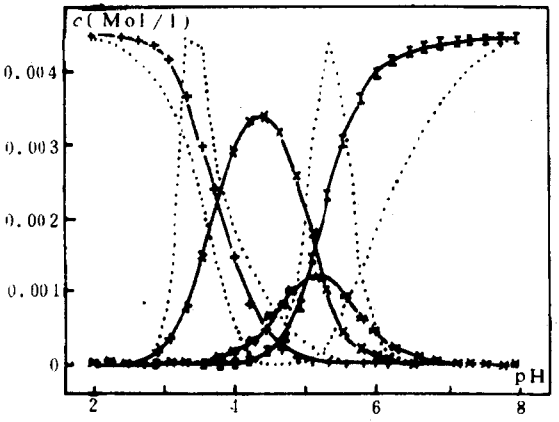


图 5.7　模型数据浓度曲线

··· 为初级 EFA 结果　— 为迭代 EFA 结果　(+ : Cu^{2+}, × : CuL^{2+}, ⋆ : $CuLH^{+}_{-1}$, I : $CuLH_{-2}$) 为质量作用定律所得结果

不用多讲，截然不同的两种分析方法所得的结果如此一致，是相当有意义的，这不仅显示了 EFA 的功能，而且在确认所选择的化学模型方面也是很有帮助的.

5.4.2　色谱中的峰分辨

对于以光谱滴定为根据的平衡研究来说，无模型的数据处理方法通常只不过作为可以替代以质量作用定律为基础的数据处理的另一种方法. 但在分析色谱数据时，情况就完全不同了，很清楚，其原因在于没有令人满意的描述色谱峰的一般模型. 当然，用高斯和洛伦兹函数或由它们的组合来作为洗提曲线的近似是可以的，但往往带有一定程度的随意性和误差倾向，因此，建立一种完全的无模型方法是极其重要的. 如前节所述，已有为数众多的方法试图解决这一问题，但都没有利用初级 EFA 所获得的“浓度窗口”这一有用信

息. M. Maeder 等人已将 EFA 成功地用于解析模仿 HPLC 的合成数据，对分辨率小到 0.2 以及对总的吸光度贡献很不相等的 3 个重叠峰，EFA 能精确地给出体系中存在的物种数以及它们的洗提曲线和吸收光谱，首先对合成的数据矩阵 (在不同的洗提时间下的 50 个光谱) 进行初级 EFA 处理，其结果示于图 5.8 中. 原始数据已显示不出有重叠峰，通过在 3 个不同洗提时间所展现出的 3 个重要的特征值很容易检测出有 3 个物种存在. 将正向 EFA(如图 5.8 中的 a) 和反向 EFA(如图 5.8 中的 c) 组合，为每一物种产生一个 "时间窗口" (如图 5.8 中的 b)，以这些窗口为基础，通过使用最小二乘法精化可获得浓度曲线和吸收光谱，如图 5.9 所示. 重叠严重的原始数据 $[Y]$ (用三维即时间、浓度和吸光度图形显示) 被解析为各物种的浓度分布 $[C]$(左边) 和光谱 $[A]$(右边)，并将它们投影到一个想象的包含原始数据的立方体的 "后墙上"，与叠加到模型数据中的随机噪音 1.00×10^{-3} 相比，吸光度的标准偏差为 1.26×10^{-3} 吸光单位，浓度的理论值和计算值之间的标准偏差为 1.2%. EFA 被用于处理 4 组分和 5 组分的严重重叠的色谱峰的分辨时也同样取得了成功.

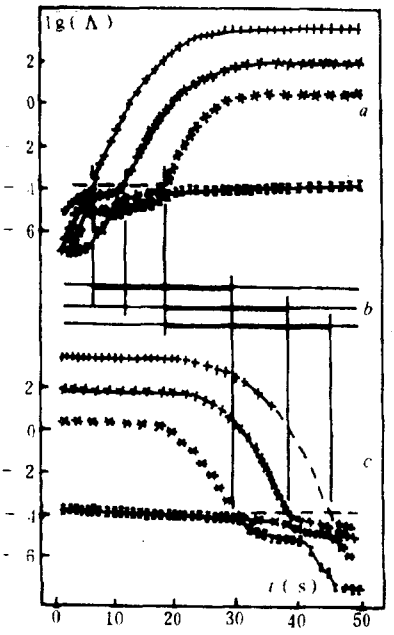


图 5.8 3 个严重重叠的色谱峰的模型数矩的 EFA 结果 特征值 $\wedge_f$ ((a) 正向 EFA) 和 $\wedge_b$ ((c) 反向 EFA) 的对数以及结果形成的时间窗口 (b) 均由相应的特征值有意义的增加所定义

许禄等提出的在重叠峰解析中 EFA 和曲线拟合联用的程序设计思想对于色谱的峰分辨工作是有参考价值的.

至此，我们可以见到，EFA 可广泛地适用于色谱中的峰分辨，而且能完成 M.F. Delaney 所描述的“曲线分辨的最终目的是能在无须对峰形、位置或种类进行假设的情况下确定重叠的色谱峰中的组分数，以及各化合物的光谱和浓度曲线”这样复杂的任务. 但是，值得一提的是，浓度分布曲线的高度和相应的吸收光谱的高度是严格遵守反比关系的，如果两者都不知道，无论使用哪种算法最后所得的结果总是模糊的，所以在上面的例子计算中，所有吸收光谱都归一化为一单位高度.

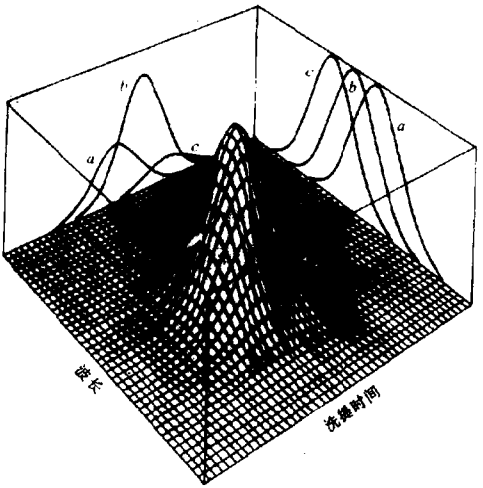


图 5.9　有 3 个组分的模型色谱，原始数据的三维表示，通过迭代 EFA 叠加成的浓度分布和吸收光谱显示在两幅“后墙”上

5.4.3　具有非理想行为的平衡混合物的分析

在许多平衡体系中，以质量作用定律为依据的模型并不能以简单可靠的方式得到应用，只要是当活度—浓度关系不能很好地确立以及主要的浓度不能直接测定时，情况就会如此. 例如，在低的和不同的离子强度下的滴定；在不同成分的混合溶剂中的实验以及在

使用 EFA 不必为溶液的 pH 值和质量作用定律的有效性所局限. 当然，在上面的情况下，以质量作用定律为基础的光谱计算基本上还是正确的，因为浓度曲线重叠并不严重，尽管如此，对于 EFA 的计算结果是可靠的和在更复杂的情况下 EFA 可能是唯一可接受的方法这两点，我们是没有理由去怀疑的.

5.4.4 产品质量控制

EFA 可用于产品质量控制及杂质检出的一些领域，尽管这不是应用的最主要目的. 关于这一点，让我们用配位化学中的例子来加以说明. 以 Cu^{2+} 与 1,4,7- 三氮癸烷为例，对于这一体系，以前曾有人用电位滴定和光谱滴定法研究过，结果说明有两种络合物 (CuL^{2+}，CuL_2^{2+}) 形成. 在广泛研究一系列的络合物的过程中，Cu^{2+} 与 1,4,7- 三氮癸烷的相互作用又被重新研究，有关数据用 EFA 方法处理.

如图 5.11 中的虚线所示，对一个含有 40 (mol)% Cu^{2+}/ 配位

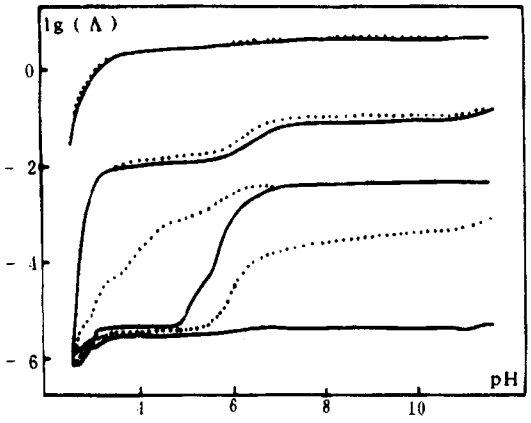


图 5.11 用 Cu^{2+} 对 1,4,7- 三氮杂环癸烷进行分光滴定的正向 EFA 作图
···: 表示采用被二聚物杂质沾污的试剂所得的结果，有 4 个有意义的特征值，
—: 表示采用经纯化的配位基所得结果，有 3 个有意义的特征值

体的滴定体系进行渐进因子分析表明，在低的 pH 范围内有 4 种吸光物质存在，其中只有 3 种 (即 Cu^{2+}， CuL^{2+} 和 CuL_2^{2+}) 是被期望的. 二聚体 $Cu_2L_2H_{-2}^{2+}$ 只在 pH 值大于 8，溶液中含有 50(mol)% Cu^{2+} 以上的情况下才能形成.

仔细观察在给定的波长处的各个滴定曲线 (吸光度对 pH)，确实发现在低的 pH 值处有一些微小的偏差，但另外一些描述其他诸如 $CuLH^{3+}$, CuL_2H^{3+} 或 $Cu(LH)_2^{4+}$ 等物种的模型经过试验并没有得到多大的成功. 尽管所研究的配位体已有完整的元素分析，而且薄层液相色谱没有显示出非同种性，但是，为慎重起见，H. Gampp 还是从环状的甲苯磺酸盐开始重新合成，并经柱色谱小心纯化，应用这一新的配位体材料，对业已用正向 EFA 处理过的这一问题得到清楚的解答，如图 5.11 的实线所示. 在对应于 CuL^{2+} 和 CuL_2^{2+} 的特征值之间的多余的那个特征值已消失，其他特征值基本上保持不变，这就有力地说明了先前所用的配位体中含有杂质. 至于原来用的产品中的杂质的性质如何，没有进行过详细的研究. 但可以设想不纯物质是 1,4,7- 三氮杂环癸烷的二聚体，当然它和不纯物的基本组成是相同的. 这样的二聚体在其他三氮杂大环化合物中的确也曾被观察到，这似乎是以 J.E. Richman 等人的甲基磺酰酯法为基础的合成中的一个有关的问题.

5.4.5 光谱数据的半定量分析及模型选择

事实证明，在许多情况下，通过 EFA 进行非常满意的定量分析是可行的，尽管如此，以特定的化学平衡体系为基础的模型方法仍在继续使用，况且计算相应的稳定常数是光谱滴定的最终目的，至少在实验能在严格控制的条件 (如温度、溶剂组成和离子强度) 下进行的研究中，情况确实如此. 即便是在这些情况下，有几种理由认为 EFA 仍不失为一种重要的分析工具，其中最显而易见的是正如以上已讨论过的那样，对于根据质量作用定律而得到的实验结果，EFA 能进行完全独立的有效检验. 如果无模型方法和有模型方法所获得

的浓度分布曲线和光谱是相同的话，那么，这就给被选择来解释数据的模型提供了具有启发性的和有力的支持. 其他理由尤其适合於复杂的体系，这时，正确选择化学模型变得模棱两可，以 EFA 为基础的最小二乘精化可能会变得复杂，在这种情况下，以初级渐进因子分析为基础的半定量分析仍然是最有用的. 下面选择 Cu^{2+} 与三氮三癸烷二酸形成的络合物为例来加以讨论，这一配位体已和其他一系列五配位基螯合剂一起用电位法和光谱法研究过. 按照一般的配位化学观点，所选体系中有两种络合物生成，即 $CuLH^{+}$ 和 CuL，它们是最稳定的型体，而且电位数据能够完满地解释这一模型. 然而，光谱数据却明显地指出这一模型是不够完善的，必须假设还有另外的物种形成. 这一发现与通常所观察到的事实是一致的，即电位滴定法的数据重现性是非常好的，但就不同模型之间辨别能力而言，它比光谱分析法要逊色一些. 在所研究的平衡体系中，额外的物种被预料是 $CuLH_2^{2+}$，也可能是 $CuLH_3^{3+}$，和 $CuL(OH)^{-}$，头两种估计在低的 pH 范围内出现，而后者则在 pH 为 9 以上出现. 如图 5.12 所示，初级 EFA 的结果是十分清楚的，总共 4 种吸光物质

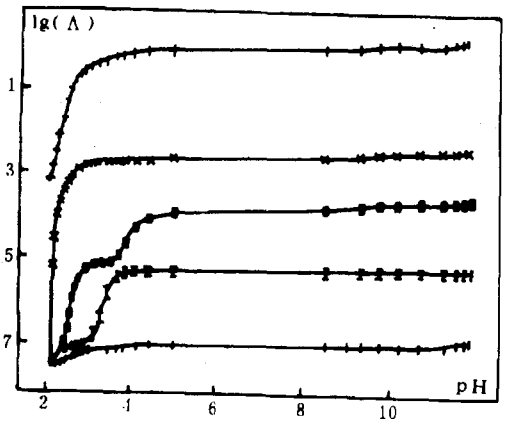


图 5.12　用 Cu^{2+} 对三氮三癸烷二酸进行分光滴定的正向 EFA 作图，显示出 pH<4 时有 4 个组分，直至 pH 12 并未有进一步的变化

均在低的 pH 范围出现，第 5 个特征值在整个实验范围内都不大. 这样，正确的模型应包括 Cu^{2+}，$CuLH_2^{2+}$, $CuLH^+$ 和 CuL，而 $CuLH_3^{3+}$ 和 $CuL(OH)^+$ 并未被检出. 的确以这个模型为基础并使用质量作用定律，光谱滴定结果可得到很好的解释，显然，EFA 的结果不仅为选择合适的模型提供算法，而且还给出各平衡常数的极好的初值. EFA 的应用优于经典的最小二乘分析，从而避免花费很多时间去试验不同的模型及其有关的任何其他准备工作，如作图估计未知的平衡常数的方法等.

上面讨论的分析体系较为简单，用通常的主成分分析法 (PCA) 也能正确地指出吸光物种数 (当然不是它们的存在范围). 在含有大于 5 或 6 种吸光物质的更复杂的体系中，PCA 就不一定能给出正确的结果. 例如，对于 Cu^{2+} 与 L- 丙氨酰胺、氨基乙酸乙酰胺以及 N,N′- 二氨基乙酰 -1,2- 乙烷二胺形成的络合物体系，每种情况下都要用有 7 种吸光物质的模型来描述. 在所有这些情况下，都不能用 PCA 的重要特征值的个数正确确定物种数，因为在 Cu^{2+} 络合物中存在宽的非结构 d-d* 跃迁. 应当承认，迭代的 EFA 在这些情况下也不能给出可靠的浓度分布曲线和光谱，同其他方法一样，EFA 有它自己的适用范围和条件. 然而，在这种复杂的情况下，初级 EFA 用作半定量分析仍然是可能的，而且对于选择合适的模型是大有帮助的，即便在吸光物种数不能直接由 PCA 确定的情况下，在给定的 pH 值处，新的物种的形成仍可通过 EFA 从它在正向和反向分析过程中对实际的特征值的影响上反映出来，这方面的内容在前面已讨论过. 可以说，第六或第七个物种即使是实际上不能产生新的较有意义的特征值，但仍然可借助已经发现的较多特征值的强烈增大来指示新的物种的形成以及估计它的存在范围.

5.4.6 在圆二向谱中的应用

目前，所有根据因子分析原理进行的自模光谱分离研究都被限制于显示出正强度值的光谱. 由于 EFA 着眼于生成浓度曲线，所以

EFA 对显示呈现出正的和负的强度的光谱方法，如圆二向谱 (CD) 和旋光色散 (ORD) 等都应该是可以应用的.

Z. Kraly 等人曾研究了 L- 苏氨烷和 L- 别苏氨酸分别和钴 (Ⅱ) 及铜 (Ⅱ) 形成的络合物的稳定常数，他们记录了这些络合物作为 pH 的函数，在 2<pH<11 范围内的 CD 谱，所得数据表明这些体系中含有 4 种类型的络合物，即 ML^+， ML_2， $(ML_2H_{-1})^{-1}$ 和 $(ML_2H_{-2})^{-2}$，其中 M 为金属离子， L 是有机配位体. 络合物的浓度曲线是由电位法测量而确定的，而后用所得的浓度信息求得各络合物的 CD 谱.

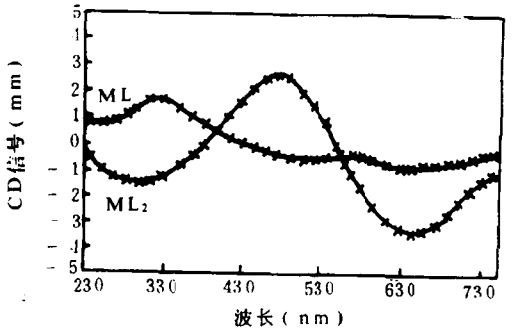


图 5.13 ML 和 ML_2 的圆二向图 (实线表示真实谱，符号点则表示 EFA 预测谱)

受到上述工作的启发，这里阐述的关于 CD 谱的研究，其目的是在原则上证明 EFA 能够在不借助电位数据和任何其他事先知道的数据的情况下，仅从 CD 混合谱便可获得存在于根本未知的络合物混合物的组分的 CD 谱以及相应的浓度分布曲线. 为了完成这一任务，先进行一个模拟研究，4 种假设的络合物 (用 ML， ML_2， ML_3 和 ML_4 表示) 的人造 CD 谱 (图 5.13 和 5.14 中的实线) 显示出正的和负的椭圆率. 这些 CD 谱跨越的波长范围为 235–745nm. 通过以 15nm 为间隔对这些谱进行数字化处理便可得到矩阵 $[E]_{\text{real}}$，这 4 种络合物的典型浓度分布曲线 (图 5.15 中实线) 被用来模拟在 pH 1 至 10.5 范围内的它们的渐进行为. 矩阵 $[C]_{\text{real}}$ 可通过在每 0.5pH 单位

间隔读取这些分布曲线得到. 这两个矩阵相乘 ($[A] = [E]_{\text{real}}[C]_{\text{real}}$) 便产生一个络合物数据矩阵. 另外, 标准偏差为 0.01 的高斯误差被加到这一矩阵中, 以便模拟更符合实际的情况.

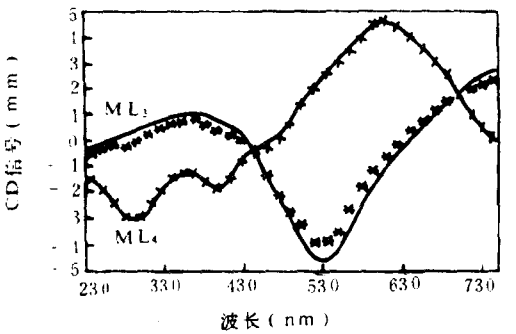


图 5.14 ML_3 和 ML_4 的圆二向图 (实线表示真实谱, 符号点则表示 EFA 预测谱)

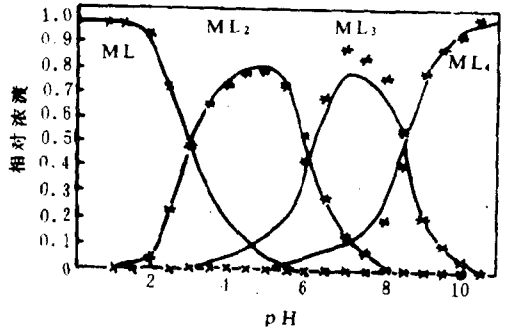


图 5.15 4 种组分的浓度分布为 pH 的函数 (实线表示真实谱, 点符号则表示 EFA 预测谱)

对构造出的 35×20 数据矩阵进行渐进因子分析, 用 Gemperline 法正确地鉴别出对应于 4 种型体中的每一种的浓度峰的最大值. 按

些结果与真实谱相当一致.

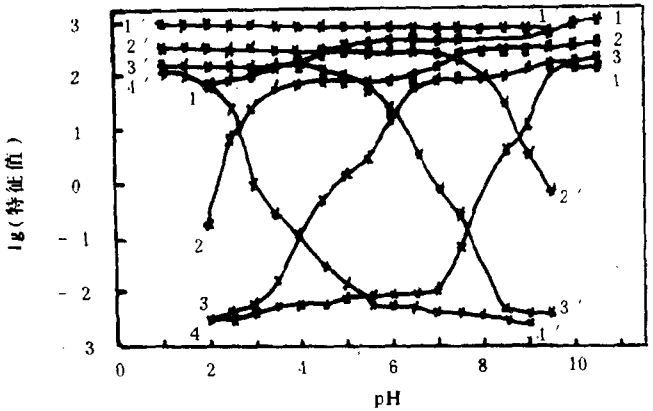


图 5.16　pH 滴定所得圆二向数据的 EFA 特征值 (无撇号者表示正向 EFA，带撇号者表示反向 EFA)

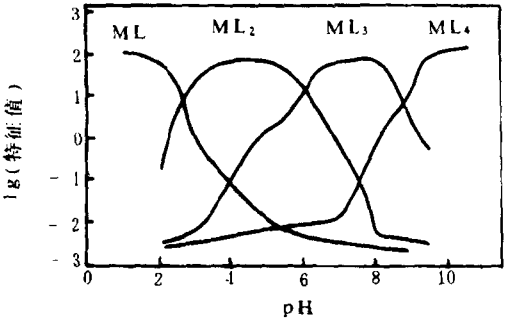


图 5.17　通过连接图 5.16 中的正向和反向 EFA 特征值图得到的未定义浓渡分布图

5.5　渐进秩消因子分析简介

B5

在第四章中，我们已谈到秩消因子分析法 (RAFA)，该法能用

于任何由遵守双线性形式的元素 $D(i,j)$ 组成的矩阵 $[D]$ 的分析

$$D(i,j)=\sum_{k=1}^{s}\beta(k)X(i,k)Y(j,k)$$

或

$$[D]=\sum_{k=1}^{s}\beta(k)[N_k]. \tag{5.27}$$

当然，应假定响应矩阵 $[N_k]$ 对要定量的物种 k 来说是已知的，即

$$N(i,j)=X(i,k)Y(j,k)$$

或

$$[N_k]=X_kY_k^T. \tag{5.28}$$

RAFA 已成功地被用来分析激发 / 发射萤光、液相色谱 / 紫外可见光谱等数据.

虽然渐进因子分析是不同于 RAFA 的另一种方法，但如果数据的排列顺序使得对于各个物种来说在 x 方向或 y 方向上具有非零响应 (即 $[N_k]$) 的不同连续范围的话，则 EFA 也可以用于分析遵守方程 (5.27) 的任意矩阵 $[D]$. 通常，在用紫外可见光谱检测的液相色谱和光谱平衡研究中的浓度分布曲线是能满足这一条件的，EFA 在这两方面的应用已取得成功. 和 RAFA 不同， EFA 不需要各个组分的响应值作为输入，事实上， X 和 Y 就是数据处理的初步结果. 因此，对于即使是纯组分的响应值不能独立测定的体系，EFA 也能应用，最明显的例子就是各物种彼此之间往往是不能分离的络合平衡体系. 在用紫外可见光谱检测的液相色谱中，虽然所感兴趣的组分的光谱常常能独立地测得，但如果使用 RAFA，必须克服一些困难，并且需要性能精良的仪器使得对不同成分和含量所组成的溶液所获得的浓度分布曲线有满意的重现性.

H. Gampp 等人把 RAFA 和 EFA 的思想相结合，提出一种新的方法，并称之为渐进秩消因子分析，简称 RAEFA. 同 RAFA 一样，

RAEFA 可用于对一未知混合物中的指定物种进行定量分析，但只须有一方向的响应是已知的 (通常是该物种的吸收光谱)，这样就避免了对色谱高重现性的要求. 在知道某些组分的吸收光谱的情况下，RAEFA 的应用也是对 EFA 的一种补充. 图 5.18 概括了总的分析方案 (图中各个符号的说明将于后面给出). 下面，简单介绍 RAEFA 的分析步骤.

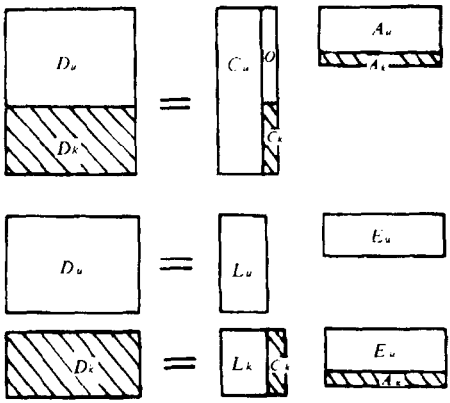


图 5.18　RAEFA 的分析过程示意图

要用 RAEFA 进行定量分析的物种的光谱 A_k 是已知的，与其相对应的未知浓度分布 C_k 通过下列步骤计算.

(1) 对完整的数据集 $[D]$(由在 w 个波长处得到的 m 个谱构成 $m \times w$ 矩阵) 进行渐进因子分析，产生一“浓度窗口”或物种存在范围. 在络合平衡研究中就自由金属离子 (最初物种) 或完全地生成的络合物 (最后物种) 而言，合适的“窗口”与对应的物种的关联并不是太重要的. 同样，在二极管阵列检测的液相色谱中，无论如何，洗脱的相对顺序总的说来是不会产生严重问题的.

(2) 用 (1) 的结果，$[D]$ 能被分解为 $[D_k]$ 和 $[D_u]$ 两部分，在此，已知的某种物种只对图 5.18 中阴影部分 $[D_k]$ 有贡献. $[D]$ 的另一部分 $[D_u]$ 仅包含未知光谱 $[A_u]$ 的线性组合，因此，$[D_u]$ 的秩为 $s-1$.

(3) 对 $[D_u]$ 进行主成分分析，得未知光谱 $[A_u]$ 的一个特征向量表象，它旋转相同的 $s-1$ 维向量空间.

(4) 通过把已知光谱 A_k 和未知光谱 $[A_u]$ 的特征向量表象 $[E_u]$ 组合可得到完整的光谱向量空间，给出一个满秩 s 的组合矩阵 $[A_c]$.

(5) 现在，光谱已知的指定物种的浓度分布 C_k 就容易计算了，它与通过对 $[A_c]$ 进行简单的线性回归 (式 (5.29)) 得到的浓度矩阵 $[C_c]$ 的某合适的列是一致的

$$[C_c] = [D_k][A_c]^T([A_c][A_c]^T)^{-1}. \tag{5.29}$$

对于 $[A_c]$ 来说，$[C_c]$ 是一个秩为 s 的混合矩阵，跨越整个浓度向量空间，它是由有明确物理意义的列向量 C_k 和其它未知浓度的抽象的特征向量表象 $[L_k]$ 组合而成.

就物种 k 而言，其分析是圆满的，不依赖浓度分布知识便已得到定量结果. 以 EFA 为基础，也可对其它光谱和浓度分布均未知的物种进行定量. 这种定量可在约化的向量空间 $[D_r]$ 中进行，$[D_r]$ 由完整数据阵减去该已知物种对数据阵的贡献的部分而得到，即

$$[D_r] = [D] - C_k A_k, \tag{5.30}$$

约化的数据矩阵 $[D_r]$ 的秩也是 $s-1$. 通过对 $[D_r]$ 进行如前所述的 EFA 分析，对未知物种进行定量分析是可能的.

RAEFA 是对一未知混合物中某给定的物种进行定量分析的一种新的有力的工具. 与 RAFA 不同，RAEFA 只需要某一些给定的吸收光谱作为输入，而不必用完整的所有物种的二维响应矩阵作为输入，它是以浓度窗口的存在为基础的，然而，这在经典的 RAFA 用于激发—发射试验的例子中是得不到的. 在由紫外可见光谱检测的液相色谱方面，RAEFA 避免需要用于获得洗提曲线的高重现性的特殊仪器设备，并且由物种间的相互作用而引起的有关柱超载或曲线变形问题也可避免. 在各个吸光物种的浓度分布不可能通过独立的实验来测定以及不可能事先了解全部响应的情况下所进行的平衡研究中，必须采用 RAEFA 来替代 RAFA，这样做，问题才有可

能获得较好的解决.

H. Gampp 等人应用这一技术在色谱峰严重重叠的 4 组分体系的研究 (借助配备有二极管阵列检测器的液相色谱技术) 以及多组分平衡体系的 pH —浓度分布计算研究中，获得了较好的结果.

6 对应因子分析

在化学的理论和应用研究中，数据的类型大多都具有多变量性质. 这是对一种化合物、一种混合物、一种物理化学性质或是其它的客体进行观察或量测而得到的性质. 要对完整的这类科学数据集进行概括和诠释，并不是用一些立面图就可以解决的.

对应因子分析 (Correspondence Factor Analysis, 简称 CFA) 是由法国的 Jean-Paul Benzécri 在 60 年代初发展起来的一种几何技术，它可被用来描述数据列联表中的行行之间、列列之间以及行列之间的双重关系. 就是说，它可被用来描述变量之间、样品之间以及变量与样品之间的双重关系. 由于数据的预处理对于样品和变量呈对称性，这就使得样品和变量可以同时被描绘在同一个图上，从而能对它们之间的相互关系做出直接的解释，尤其是，能够将变量与那些对于变量来说特别有意义的样品相关联.

这种技术在化学基础理论及应用研究中有其独特的用途，它能帮助化学家根据大量的数据做出正确的判断.

6.1 一些典型问题介绍

石油化学的发展吸引着色谱学家不断地去对烃类在气 - 液相色谱以及气 - 固相色谱中的 Kovats 保留指数进行系统的测定. 对于这些数据的概括以及对它们的物理化学意义的解释当然也一直是色谱学家们所关注的问题.

如果相类似的特效或非特效相互作用支配着反应和保留作用，则同分异构烃类的色谱性质可同它们的反应性关联. 例如，烃类在氧化硅或氧化铝上的吸附肯定是在这些材料上的气相色谱学和裂解

过程的催化的一个基本方面，在阮来镍上的非均相催化或在钨体系上的均相催化中的烯烃类的反应性主要取决于空间位阻和变形．另一方面，在石墨化碳黑上的保留作用则是由于非特效相互作用而造成．从这些固定相上获得的数据使得测定烷烃类和烯烃类的空间位阻成为可能．

烯烃类碳 - 碳双键的亲电子反应性受到极性和电荷转移络合物的生成的影响．对于烯烃类，已经清楚的显示出极性效应和空间效应之间的竞争，但是，分辨这两种效应却是困难的．通过在气 - 液相色谱中采用不同极性的固定相，有可能在实践上从非特效性相互作用转变成高特效性相互作用，遗憾的是，随着增加固定相的极性，烷基链的加长或支化将会减小烯烃双键和固定相之间的特效相互作用．为了避免这种局限，为了方便这方面的色谱学内容的研究，必须充分利用特效相互作用的可能性，这样做，对于弄清楚这些特效相互作用的物理化学机理也是有帮助的．

在不同金属离子形式的大网目离子交换树脂上的气 - 固相色谱法看起来是可以达到上述目的的手段，根据所用的离子的不同，它可在烃类中提供好的选择性．轻度磺化过的多孔高分子树脂具有高的柱操作效率可以改变磺化程度，随着这种改变，对于在填充料上与金属平衡离子生成络合物的化合物来说，特效性的范围也改变．

为了能达到上面所讲到的目的，就必须系统地研究阳离子性质以及各种烃类的结构与性质对这些溶质的保留作用的影响．为了对所涉及的特效相互作用中的主要趋势做出较准确的描述，有人在研究中采用了 40 种烃类 (包括烷类、环烷类、烯类、支链烯类、环烯类以及芳香类化合物) 和 9 种阳离子．

溶质和固定相之间存在的相互作用使色谱的保留现象变得复杂，上述研究的过程中共获得 300 多个保留指数数据．很清楚，要对这么多的数据进行解释并不是件容易的事情．借助“对应因子分析”技术，可通过揭示溶质 - 固定相相互作用而增大从测量结果所提炼出的信息量，可以同时对溶质和固定相进行分类，可通过选择性中

的主要趋势来了解系统的复杂性. 这种主要的趋势是由分子的性质和结构 (功能团性质的影响, 分子的取代程度, 支链化或烷基链加长的影响和极性等) 所造成的.

又如, 有人测定了 47 种芳丙烯酰芳烃类化合物在 5 种极性不同的固定相 (Zorbax C_8, Zorbax ODS, Zorbax NH_2, Lichrospher 100 DIOL 和 MicroPak CN) 上分离的容量因子 (共 235 个数据). 这些烃类包括 E-s-cis 和 Z-s-cis 两类同分异构体. 采用对应因子分析技术可以对这些烃类在所研究的色谱系统中的色谱行为进行比较, 也可讨论对于已分离出的各个子类的烃类的色谱系统的特效性以及位置的和构型的同分异构现象对分离作用的影响.

在类似于上述的科学研究工作中, 化学工作者往往都会面临着如何从手头所获得的众多的数据中提取出更多有用的信息的问题, 也就是如何来对大量的数据进行科学的概括和解释的问题. 在这种情况下, 对应因子分析技术更能显示其独特的优点.

6.2 对应因子分析的数学描述

考虑到本书编写的目的和对象, 在这里我们只介绍对应因子分析的一般数学处理过程. 对于过程中的各个步骤不作严格的数学证明而只予以适当的解释, 以求读者能明了易懂并尽快应用于自己的科学研究和生产实践中.

假设待进行对应因子分析的数据矩阵 $[D]_{I\times J}$ 是一个非负数据矩阵且它的行加和、列加和均为非零, 即 $[D]_{I\times J}\equiv[d_{ij}]$, $d_{ij}\geq 0$. 对 $[D]_{I\times J}$ 进行对应因子分析的全过程主要包括两大部分: ①数据预处理; ②降维处理. 下面分两个小节来分别介绍这两个部分.

6.2.1 数据预处理

先定义原始数据矩阵的对应矩阵 $[P]$, 它的元素是 $[D]$ 中的对

应元素被 $[D]$ 的总和去除

$$[P] \equiv (1/t)[D], \tag{6.1}$$

式中 $t = \sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij}$，矩阵 $[P]$ 的行加和向量和列加和向量分别用 $\mathbf{r}$ 和 $\mathbf{c}$ 表示

$$\left.\begin{array}{l} \mathbf{r} \equiv [P]\mathbf{1} \ \ (\mathbf{1} \equiv [1 \cdots 1]^T, \text{ 是一个 J 元向量}, \\ \mathbf{c} \equiv [P]^T\mathbf{1} \ \ (\mathbf{1} \equiv [1 \cdots 1]^T, \text{ 是一个 I 元向量}. \end{array}\right\} \tag{6.2}$$

其中 $r_i > 0\ (i = 1, 2, \cdots, I)$, $c_j > 0\ (j = 1, 2, \cdots, J)$, 由这些加和值所组成的对角阵分别用 $[D_r]$ 和 $[D_c]$ 表示

$$\left.\begin{array}{l} [D_r] \equiv \mathrm{diag}(\mathbf{r}), \\ [D_c] \equiv \mathrm{diag}(\mathbf{c}). \end{array}\right\} \tag{6.3}$$

矩阵 $[P]$ 的加和等于 1，把 $[D]$ 看作是一个列联表，则 $[P]$ 可被看作是一个 $I \times J$ 概率密度矩阵，$\mathbf{r}$ 和 $\mathbf{c}$ 便是边缘密度.

定义矩阵 $[P]$(也就是矩阵 $[D]$) 的行分布和列分布分别为 $[P]$(或是 $[D]$) 的行向量和列向量被它们各自的加和去除，设行分布和列分布的矩阵分别为 $[R]$ 和 $[C]$

$$[R] \equiv [D_r]^{-1}[P] \equiv \begin{bmatrix} \tilde{R}_1^T \\ \cdots \\ \tilde{R}_i^T \end{bmatrix}, \ [C] \equiv [D_c]^{-1}[P]^T \equiv \begin{bmatrix} \tilde{C}_1^T \\ \cdots \\ \tilde{C}_j^T \end{bmatrix}. \tag{6.4}$$

在这里，行分布 $\tilde{R}_i (i = 1, \cdots, I)$ 和列分布 $\tilde{C}_j (j = 1, \cdots, J)$ 分别以 $[R]$ 和 $[C]$ 的行形式来书写. 实际上这些分布正好等于 $[D]$ 的行和列被它们各自的加和去除. 同理，$\mathbf{r}$ 和 $\mathbf{c}$ 也就等于 $[D]$ 的行加和、列加和被 t 去除. 在实际处理中，用 $[P]$ 要比用 $[D]$ 更方便一些，因为对应因子分析结果只同数据的相对值有关.

行分布矩阵 $[R]$ 在 J 维加权欧氏空间中定义了一个点群，这个点群中包含有 I 个点，每个点代表在 J 维空间中的一个行分布 $\tilde{R}_i$，

此时，加权空间的维数均由 $\mathbf{c}$ 中各元素的倒数 (即 $[D_c]^{-1}$) 来定义，同理，列分布矩阵 $[C]$ 在 I 维加权欧氏空间定义了一个点群，该点群包含有 J 个点，每个点代表在 I 维空间中一个列分布 $\tilde{C}_j$，此时，加权空间的维数均由 $\mathbf{r}$ 中各元素的倒数 (即 $[D_r]^{-1}$) 来定义. 很明显，行分布的点群和列分布的点群在各自的空间中的重心分别是 $\mathbf{r}$ 和 $\mathbf{c}$. 每一个点群的总空间变差可用它们的总惯量值来定量. 总惯量值就是从各个点至它们相应的重心的距离平方的加权加和. 行点的总惯量 (用 $\mathrm{in}(I)$ 表示) 为

$$\mathrm{in}(I) = \sum r_i(\tilde{R}_i - \mathbf{c})^T[D_c]^{-1}(\tilde{R}_i - \mathbf{c}), \tag{6.5}$$

即

$$\mathrm{in}(I) = \mathrm{trace}([D_r]([R] - \mathbf{1c}^T)[D_c]^{-1}([R] - \mathbf{1c}^T)^T). \tag{6.6}$$

列点的总惯量 (用 $\mathrm{in}(J)$ 表示) 为

$$\mathrm{in}(J) = \sum c_j(\tilde{C}_j - \mathbf{r})^T[D_r]^{-1}(\tilde{C}_j - \mathbf{r}), \tag{6.7}$$

即

$$\mathrm{in}(J) = \mathrm{trace}([D_c]([C] - \mathbf{1r}^T)[D_r]^{-1}([C] - \mathbf{1r}^T)^T) \tag{6.8}$$

6.2.2 降维处理

行点群和列点群的各自的低维 (k^* 维) 子空间 (这些子空间就距离的平方的加权加和而言，是同这些点最接近的) 分别被用 $[P] - \mathbf{rc}^T$ 的 k^* 个右的和左的广义奇异向量来定义 (这些奇异向量分别以 $[D_r]^{-1}$ 和 $[D_c]^{-1}$ 为度量，且对应于 k^* 个最大的特征值). 换句话说，右奇异向量和左奇异向量分别定义行点群和列点群的主要因子轴. 设 $[P] - \mathbf{rc}^T$ 的广义奇异值分解为

$$[P] - \mathbf{rc}^T = [A][D_\mu][B]^T, \tag{6.9}$$

式中

$$[A]^T[D_r]^{-1}[A] = [B]^T[D_c]^{-1}[B] = [I], \tag{6.10}$$

式中

$$[\tilde{M}]^T[D_c][\tilde{M}] = [\tilde{N}]^T[D_r]^{-1}[\tilde{N}] = [I], \tag{6.21}$$

最后可得到列分布在由 $[\tilde{N}]$ 的列所定义的空间中的投影 $[G]$ 为

$$[G] = [\tilde{M}][D_\mu], \tag{6.22}$$

由此可知，在最佳的 k^* 维子空间中行分布和列分布的坐标就分别为

$$[F_{(k^*)}] = [\tilde{N}_{(k^*)}][D_{\mu(k^*)}], \tag{6.23}$$

$$[G_{(k^*)}] = [\tilde{M}_{(k^*)}][D_{\mu(k^*)}], \tag{6.24}$$

实际上，矩阵 $[G]$ 和 $[\tilde{N}]$ 与矩阵 $[M]$ 和 $[F]$ 之间存在着以下的转换关系

$$[G] = [D_c]^{-1}[M][D_\mu], \tag{6.25}$$

$$[\tilde{N}] = [D_r][F][D_\mu]^{-1}, \tag{6.26}$$

所以，在实际计算中，完成行分布后，采用得到的 $[M]$ 和 $[D_\mu]$ 按式 (6.25) 即可算得列分布的坐标.

到此为止，我们已经能够做到将大的数据表转化成概括性的和易于解释的图示从而使得存在于多维数据表中的原来的信息尽可能好地被保留在一个二维图中.

不过，必须注意，虽然这样的二维图是“最佳”的平面图象，但它毕竟只是多维空间在某一平面上的投影，在某些情况下，在二维图中相邻近的点事实上在多维空间中却相去较远，因此，有必要引进“贡献”的概念.

每一个点群的总惯量可沿着各个主轴和围绕着各个点做相似而对称的分解 (同方差分解相类似)，结果如表 6.1 所示.

表 6.1 惯量的分解

	轴					
		1	2	⋯	K	总计
行	1	$r_1 f_{11}^2$	$r_1 f_{12}^2$	⋯	$r_1 f_{1k}^2$	$r_1 \sum_{k=1}^{K} f_{1k}^2$
	2	$r_2 f_{21}^2$	$r_2 f_{22}^2$	⋯	$r_2 f_{2k}^2$	$r_2 \sum_{k=1}^{K} f_{2k}^2$
	⋯	⋯	⋯	⋯	⋯	⋯
	I	$r_I f_{I1}^2$	$r_I f_{I2}^2$	⋯	$r_I f_{Ik}^2$	$r_I \sum_{k=1}^{K} f_{Ik}^2$
	总计	$\lambda_1 \equiv \mu_1^2$	$\lambda_2 \equiv \mu_2^2$	⋯	$\lambda_K \equiv \mu_K^2$	in(I)=in(J)
列	1	$c_1 g_{11}^2$	$c_1 g_{12}^2$	⋯	$c_1 g_{1k}^2$	$c_1 \sum_{k=1}^{K} g_{1k}^2$
	2	$c_2 g_{21}^2$	$c_2 g_{22}^2$	⋯	$c_2 g_{2k}^2$	$c_2 \sum_{k=1}^{K} g_{2k}^2$
	⋯	⋯	⋯	⋯	⋯	⋯
	J	$c_J g_{J1}^2$	$c_J g_{J2}^2$	⋯	$c_J g_{Jk}^2$	$c_J \sum_{k=1}^{K} g_{Jk}^2$

我们将表 6.1 中的列分别叫做行和列对一个轴的惯量的贡献，可以将每一个这些贡献表示成各自的轴的比例部分，以便解释该轴本身，由于这些贡献受到每一个点的质量 (实际上是加权) 的影响，故这类贡献被叫做 “绝对贡献”，表中的每一行都包含有各个轴对相应的分布点的惯量的贡献. 当然，也可以将这些贡献表示成各个点的惯量的比例部分以便解释该分布点在各轴上的 “好” 的程度. 由于每一个点的质量已被约去，故这类贡献就被叫做 “相对贡献”.

行分布点 i 对轴 k 的惯量的绝对贡献为

$$AB_i^{(k)} = \frac{r_i f_{ik}^2}{\lambda_k}; \tag{6.27}$$

轴 k 对行分布点 i 的相对贡献为

$$RB_k^{(i)} = \frac{f_{ik}^2}{\sum_{k=1}^{K} f_{ik}^2}; \tag{6.28}$$

列分布点 j 对轴 k 的惯量的绝对贡献为

$$AB_j^{(k)} = \frac{C_j g_{ik}^2}{\lambda_k}; \tag{6.29}$$

轴 k 对列分布点 j 的相对贡献为

$$RB_k^{(j)} = \frac{g_{jk}^2}{\sum_{k=1}^{K} g_{jk}^2}. \tag{6.30}$$

6.3 对应因子分析的数据实例

为帮助读者更直观地理解节 6.2 的内容并检验自己所编对应因子分析程序的正确性，这里完整地介绍用对应因子分析技术处理一个非负数据矩阵的全过程. 很明显，为了达到上述目的，给出一些必要的中间过程结果是有益的.

设 $[D]$ 为 5×4 数据矩阵，为便于说明，在这里用列联表形式来表示 $[D]$ 及其对应矩阵 $[P]$

$[D]$ 矩阵

列 (J) / 行 (I)	1	2	3	4	$\sum^{J}$
1	4	2	3	2	11
2	4	3	7	4	18
3	25	10	12	4	51
4	18	24	33	13	88
5	10	6	7	2	25
$\sum^{I}$	61	45	62	25	193($=t$)

$$[R]-1\mathbf{c}^T=\begin{bmatrix}0.048 & -0.051 & -0.049 & 0.052\\ -0.094 & -0.066 & 0.068 & 0.093\\ 0.174 & -0.037 & -0.086 & -0.051\\ -0.111 & 0.040 & 0.054 & 0.018\\ 0.084 & 0.007 & -0.041 & -0.050\end{bmatrix},$$

按式 (6.13) 算得的矩阵 $[S]$

$$[S]=\begin{bmatrix}0.0202 & -0.0254 & -0.0204 & 0.0347\\ -0.0510 & -0.0421 & 0.0364 & 0.0786\\ 0.1592 & -0.0395 & -0.0780 & -0.0730\\ -0.1339 & 0.0553 & 0.0640 & 0.0341\\ 0.0537 & 0.0510 & -0.0262 & -0.0495\end{bmatrix}.$$

求 $[S]^T[S]$ 的特征值和特征向量得

$$\begin{aligned}\lambda_1&=7.475908\times10^{-2},\quad \text{占总惯量的}\quad 87.756\%;\\ \lambda_2&=1.001718\times10^{-2},\quad \text{占总惯量的}\quad 11.759\%;\\ \lambda_3&=4.135744\times10^{-4},\quad \text{占总惯量的}\quad 0.4855\%.\end{aligned}$$

λ_1 和 λ_2 的累加已占总惯量的 99.5145%，所以，实际上只需考虑矩阵 $[S]$ 的 2 秩近似就可以了 (为了读者对照方便，我们在此仍给出 3 秩近似的结果)，即为

$$[S_{(3)}]=[U_{(3)}][D_{\mu(3)}][V_{(3)}]^T,$$

其中
$$[U_{(3)}]=\begin{bmatrix}0.0574 & 0.4621 & 0.8333\\ -0.2892 & 0.7424 & -0.5061\\ 0.7155 & 0.0548 & -0.1303\\ -0.5753 & -0.3896 & 0.1098\\ 0.2647 & -0.2838 & -0.1430\end{bmatrix}$$

$$[D_{\mu(3)}]=\begin{bmatrix}0.27342 & & 0\\ & 0.10008 & \\ 0 & & 0.02034\end{bmatrix},$$

$$[V_{(3)}]^T=\begin{bmatrix}0.8087 & -0.1756 & -0.4070 & -0.3867\\ 0.1713 & -0.6806 & -0.0417 & 0.7112\\ -0.0246 & 0.5223 & -0.7151 & 0.4639\end{bmatrix},$$

按式 (6.15)， (6.16) 分别算得

$$[N_{(3)}] = \begin{bmatrix} 0.2405 & 1.9357 & 3.4903 \\ -0.9471 & 2.4310 & -1.6574 \\ 1.3920 & 0.1065 & -0.2535 \\ -0.8520 & -0.5769 & 0.1625 \\ 0.7355 & -0.7884 & -0.3974 \end{bmatrix},$$

$$[M_{(3)}] = \begin{bmatrix} 0.4546 & 0.0963 & -0.0138 \\ -0.0848 & -0.3286 & 0.2522 \\ -0.2307 & -0.0236 & -0.4053 \\ -0.1392 & 0.2560 & 0.1670 \end{bmatrix}.$$

然后，再按式 (6.19)， (6.25) 分别可算出 $[F]$ 和 $[G]$

$$[F] = \begin{bmatrix} 0.0658 & 0.1937 & 0.0710 \\ -0.2590 & 0.2433 & -0.0337 \\ 0.3806 & 0.0107 & -0.0052 \\ -0.2330 & -0.0577 & 0.0033 \\ 0.2011 & -0.0789 & -0.0081 \end{bmatrix},$$

$$[G] = \begin{bmatrix} 0.3933 & 0.0305 & -0.0009 \\ -0.0995 & -0.1411 & 0.0220 \\ -0.1963 & -0.0074 & -0.0257 \\ -0.2938 & 0.1978 & 0.0262 \end{bmatrix}.$$

最后按式 (6.27)—(6.30) 算出各行分布点，列分布点对各轴的贡献 (即绝对贡献) 以及各轴对各分布点的相对贡献，全部结果列于表 6.2 中.

将 $[F]$ 和 $[G]$ 中的前两列数值分别为行分布点和列分布点的坐标，将它们描绘在由轴 1 和轴 2 组成的平面上即可得到这些分布点的图示 (见图 6.1).

表 6.2　绝对贡献 (AB) 和相对贡献 (RB) 的结果 (%)

分布点 \ 轴		轴 1		轴 2		轴 3	
		AB	RB	AB	RB	AB	RB
行点	1	0.330	9.223	21.356	80.034	69.433	10.743
	2	8.366	52.640	55.115	46.468	25.619	0.892
	3	51.201	99.903	0.300	0.078	1.698	0.018
	4	33.097	94.193	15.177	5.788	1.205	0.019
	5	7.006	86.534	8.052	13.326	2.045	0.140
列点	1	65.400	99.402	2.934	0.597	0.061	0.0005
	2	3.085	32.672	46.317	65.729	27.282	1.598
	3	16.562	98.185	0.174	0.138	51.140	1.677
	4	14.954	68.440	50.576	31.015	21.518	0.545

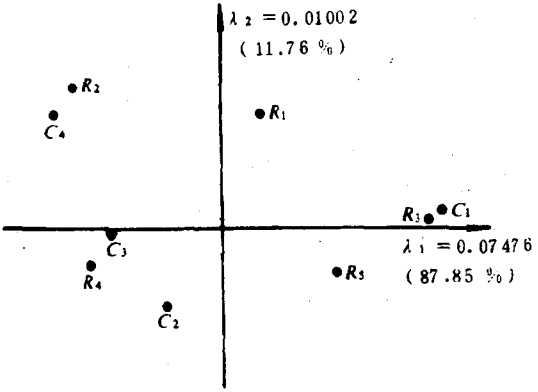


图 6.1　行分布和列分布的图示

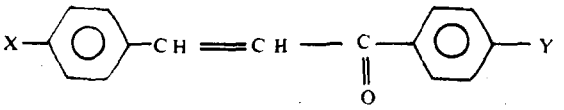
$R_1 - R_5$ 表示行分布点，$C_1 - C_4$ 表示列分布点

6.4　对应因子分析的应用

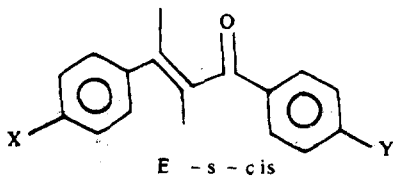
为了评估对在分析或物理化学应用中的选择性有影响的主要因素，为了研究通过色谱数据来模化系列药物的活性的可能性，B. Walczak 等人根据对应因子分析技术设计了这样的实验：用一系列

芳丙烯酰芳烃类作为模型化合物并被采用反相高效液相色谱 (RP-HPLC) 来进以分析，采用非极性化学键合固定相. 通过实验，获得了对苯环上 4- 和 4′- 位置的取代基的憎水性的影响的定量认识，发现对位置的同分异构作用的憎水灵敏性取决于构型的同分异构作用. 在药物化学中，通过细胞膜的药物输送可用 Hansch 参数或是 RP-HPLC 数据来进行模化；而在药物作用中，与受体位置的特效相互作用也显得非常重要.

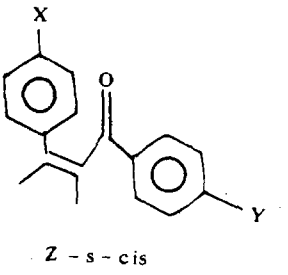
为了补充从 RP-HPLC 所获得的结果， B. Walczak 等又对正相高效液相色谱 (NP-HPLC) 中潜在的特效相互作用进行系统研究. 为了确定羰基、苯环和 4- 或 4′- 取代基的相对贡献，一系列相类似的芳丙烯酰芳烃



(X 和 Y 表示不同的原子或基团) 在极性不同的固定相上予以分离. 此外，为了便于了解经选择过的芳丙烯酰芳烃组 (E-s-cis 和 Z-s-cis) 在不同极性的固定相上的色谱系统中的分离情况，故在实验中采用了 5 种极性不同的固定相： Zorbax ODS, Zorbax C_8, Zorbax NH_2, LiChrospher 100 DIOL 和 MicroPak CN. 实验测得的各种芳丙烯酰芳烃类化合物在不同柱上的容量因子的数据列于表 6.3 中.



E - s - cis



Z - s - cis

表 6.3　47 种 E-s-cis 和 Z-s-cis 芳丙烯酰芳烃在不同柱上被分离的容量因子

No.	芳丙烯酰芳烃类 X – Y	NH_2	DIOL	CN	ODS	C_8
1	H-CF_3	3.39	1.14	2.99	0.22	0.29
2	H-tBu	3.49	1.21	3.22	0.37	0.41
3	H-iPr	3.52	1.22	3.31	0.41	0.43
4	H-H	4.04	1.37	3.85	0.67	0.52
5	F-H	4.82	1.67	4.47	0.78	0.55
6	H-F	4.03	1.38	3.7	0.61	0.48
7	H-Et	3.84	1.32	3.56	0.51	0.47
8	H-Me	4.25	1.43	3.88	0.67	0.52
9	F-Me	4.93	1.69	4.45	0.81	0.53
10	F-F	4.6	1.76	4.46	0.71	0.52
11	Me-ϕ	6.98	2.25	5.98	0.75	0.5
12	MeO-Me	11.94	3.31	9.29	1.88	1.04
13	Me-MeO	12.07	3.35	9.22	2.03	1.08
14	F-MeO	14.29	4.06	10.74	2.09	1.16
15	H-NO_2	11.29	3.4	9.99	1.17	0.7
16	MeO-ϕ	19.84	5.19	14.32	2.23	1.03
17	F-NO_2	15.4	4.57	13.14	1.4	0.75
18	NO_2-Me	16.95	14.69	13.85	1.96	0.88
19	NO_2-H	17.43	4.93	14.99	1.85	0.98
20	MeO-MeO	32.38	7.51	21.59	6.6	2.05
21	MeO-NO_2	22.51	6.82	19.91	2.95	1.26
22	NO_2-F	23.23	5.87	16.94	2	0.96
23	$N(Me)_2$-NO_2	42.83	9.2	29.35	5.33	1.74
24	NO_2-MeO	56.67	11.21	34.02	5.33	2
1*	H-CF_3*	3.23	0.92	2.71	0.36	0.29
2*	H-tBu*	3.49	0.96	3.01	0.22	0.41
3*	H-iPr*	3.52	0.97	3.08	0.18	0.43
4*	H-H*	3.72	1.06	3.34	0.57	0.52
5*	F-H*	3.99	1.12	3.23	0.66	0.55
6*	H-F*	4.03	1.07	3.24	0.61	0.48
7*	H-Et*	3.84	1.04	3.26	0.44	0.47
8*	H-Me*	4.05	1.1	3.53	0.57	0.52
9*	F-Me*	4.13	1.12	3.22	0.66	0.53

续表 6.3

10*	F-F*	4.6	1.24	3.43	0.71	0.52
11*	Me-ϕ*	6.36	1.63	5.98	0.59	0.5
12*	MeO-Me*	8.46	1.94	6.17	1.09	0.72
13*	Me-MeO*	10.32	2.31	7.31	1.49	0.91
14*	F-MeO*	11.29	2.52	7.38	1.59	1
15*	H-NO_2*	10.69	2.53	8.15	1.29	0.72
16*	MeO-ϕ*	14.21	3.03	9.49	1.36	0.74
17*	F-NO_2*	16.65	3.39	10.39	1.94	1.02
18*	NO_2-Me*	16.95	3.4	10.83	2.47	1.17
19*	NO_2-H*	17.43	3.64	11.82	2.26	1.28
20*	MeO-MeO*	22.47	4.29	13.9	3.85	1.46
21*	MeO-NO_2*	30.92	4.05	12.78	2.12	1.03
22*	NO_2-F*	27.08	4.54	14.16	3.32	1.63
24*	NO_2-MeO*	49.67	7.4	24.47	5.86	2.47

柱：NH_2 Zorbax NH_2, DIOL LiChrospher 100 DIOL, CN MicroPak CN, ODS Zorbax ODS, C_8 Zorbax C_8

流动相：庚烷 – 四氢呋喃 (97:3)

检测：UV 在 280 nm, ⋆ 表示 Z-s-cis 芳丙烯酰芳烃类

缩写：tBu 特丁基，iPr 异丙基，Et 乙基，Me 甲基，ϕ 苯基

为了深入地了解所研究的色谱体系的特性，对表 6.3 中所列的原始数据进行对应因子分析，结果列于表 6.4 以及示于图 6.2，6.3 之中.

表 6.4 47 种芳丙烯酰芳烃类的对应因子分析：5 种色谱系统对惯量主轴的贡献

固定相	轴 1	轴 2	轴 3
Zorbax NH_2	43.3	1.2	6.9
Lichrosper 100 DIOL	17.1	2.6	6.6
MicroPak CN	22.8	7.0	2.2
Zorbax ODS	1.3	60.8	31.6
Zorbax C_8	15.6	28.3	52.7

平面上投影的邻近证明了它们在芳丙烯酰芳烃类分离方面的相似性(即，可观察到在这两种体系之间的容量因子的比例性).

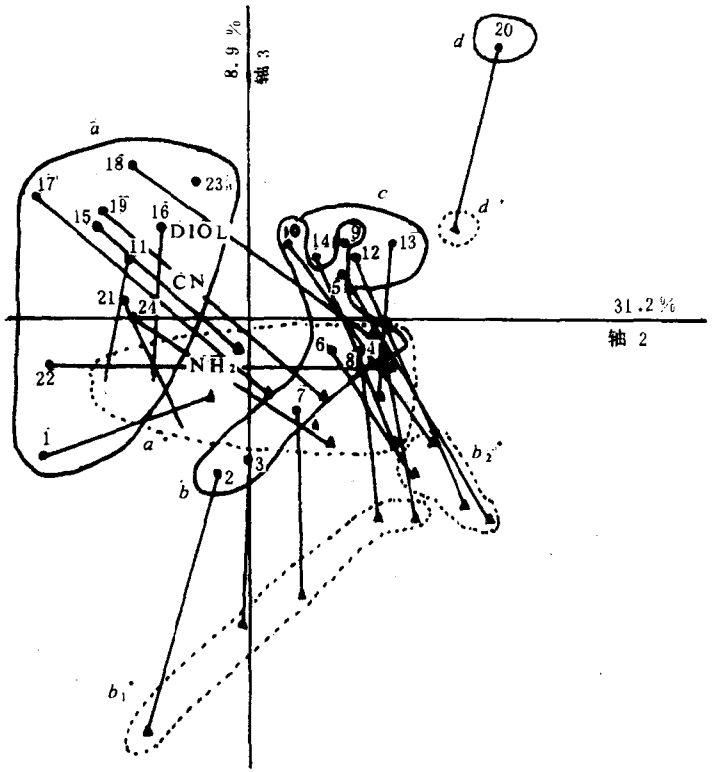


图 6.3　47 种芳丙烯酰芳烃类在惯量主轴 2 和 3 所组成的平面上的分布

现在，让我们来观察 E-s-cis 芳丙烯酰芳烃类在两个投影平面上的排布情况. 从图 6.2 和 6.3 可以明显地见到：所研究的芳丙烯酰类形成 4 个聚类，它们分别含有下列基团： (a) NO_2 ， ϕ 和 CF_3 ； (b) 烷基和 F ； (c)MeO ； 和 (d) MeO—MeO. 这些聚类主要由轴 1 和轴 2 来决定. 此外，轴 3 可以区分各子类之中的各种化合物. 在 b 类中，含 F-F, F--Me 和 F-H 的化合物在轴 3 上具有正的坐标，

其余的化合物则有负的坐标.

根据上面所指出的 E-s-cis 芳丙烯酰芳烃类化合物的聚类排布情况以及在所讨论的投影上的色谱体系的情况，可以得出这样的结论： E-s-cis 芳丙烯酰芳烃类的这种分类是由 NO_2-， ϕ- 和 CF_3- 取代的芳丙烯酰芳烃类和极性固定相之间的相互作用、 MeO- 取代的芳丙烯酰芳烃类和 ODS 固定相之间的相互作用以及烷基和 F- 取代的芳丙烯酰芳烃类和 C_8 固定相之间的相互作用所造成的.

Z-s-cis 芳丙烯酰芳烃类在两个平面上的投影的相对分布情况 (图 6.2 和 6.3 中的 a^*, b_1^*, b_2^*, c^* 和 d^* 聚类) 与 E-s-cis 芳丙烯酰芳烃类的 (图中的 a, b, c, d 聚类) 完全不相同. a 类 (含有 NO_2-, ϕ- 和 CF_3- 取代芳丙烯酰芳烃类) 同 c 类 (含有 MeO- 取代芳丙烯酰芳烃类) 有很好的分离，而 a^* 类同 c^* 类都挨在一起；含 F- 和烷基 - 取代的芳丙烯酰芳烃类却同时形成两个分离的小组 ——b_1^* 和 b_2^*. 此外， b_1^* 组沿轴 3 方向同其他的组分离开，轴 3 也反映了 MeO-MeO* 芳丙烯酰芳烃类在 ODS 固定相上的特效行为. E-s-cis 芳丙烯酰芳烃类在两个投影平面上的不同分布图案说明构型的同分异构现象对芳丙烯酰芳烃类分离的大的、非均匀的影响. 这种非均匀的影响取决于芳丙烯酰芳烃取代基的化学性质.

下面再借助因子分析结果来讨论位置不同的同分异构体. 挑选 10 对位置不同的同分异构体并参照图 6.3 的结果简化描绘成图 6.4 和图 6.5. 这 10 对位置不同的同分异构体不仅保留作用不相同，而且，它们对色谱条件变化的敏感性也不相同.

沿着轴 1 和轴 3，可以观察到差异最大的是 NO_2-F 和 F-NO_2，但对于 NO_2-F* 和 F-NO_2^* 却没有这种现象， NO_2-F* 和 F-NO_2^* 沿着轴 2 方向上的差异是有特点的. 对于取代同分异构体，另外一种规律性是典型的： Z-s-cis 构型的取代同分异构体在各轴上位置的差异比 E-s-cis 构型的要大，这种差异在轴 2 上总是最大的. 比较沿着某一给定的轴的移动方向可以对化合物的性质变化提供另外的信息，即 NO_2 或 F 基团从位置 4′ 转移至位置 4 时在轴 2 上引起的移动方

对位置不同的或构型不同的同分异构体的最佳分离的目的.

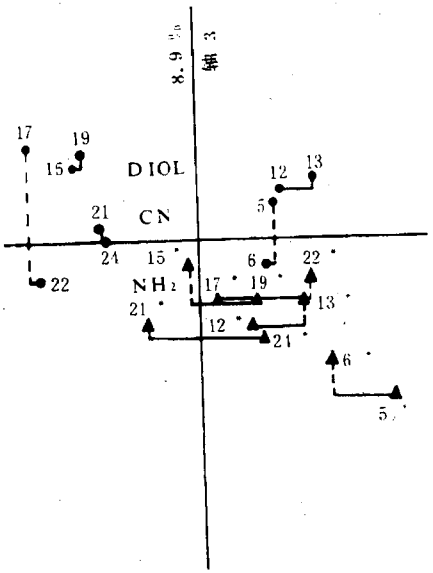


图 6.5　图 6.3 的简化图

另一个较典型的例子是 R.F. Hirsch 等人对烃类在阳离子交换树脂上的气 - 固色谱选择性趋势的研究，他们提供了 40 种烃类化合物在 9 种不同的离子交换树脂固定相上的 Kovats 指数数据. 最后对 33 种烃类在 7 种不同固定相上的 Kovats 指数构成的数据矩阵进行对应因子分析，结果指出 Ag^+， Ni^{2+}， Zn^{2+} 和 Cd^{2+} 离子型的离子交换树脂对不饱和烃类产生不同的选择性，这种现象取决于烃类的性质、取代的程度和形成 π 键的能力. 对于较大的阳离子，如 Ag^+，这种选择性的位阻现象就会出现，相对于非磺化共聚物基体来说，H^+， K^+， Na^+，和 Tl^+ 离子型的交换树脂实际上是没有选择性的.

P. Robert 等人在通过对磨碎的硬质小麦产品的中红外漫反射谱进行多元统计分析借以对产品的纯度进行定量表征的研究中，成功

地采用了对应因子分析技术来达到选定小麦中主要组分 (如淀粉、纤维素和非纤维细胞壁成分等) 的特征数. 他们还在对牛奶的近红外谱的研究中, 成功地采用主因子分析技术克服由于水的大吸收以及牛奶的散射颗粒脂肪小球使光谱变型的障碍, 从而选定了脂肪和蛋白质的特征波数.

M. Feinberg 在他的综述性文章中介绍了对应因子分析在他们实验室中的应用情况. 他们在取样设计的准备工作中, 在昆虫行为选定性的描述中以及在环境污染的评价中都普遍采用了对应因子分析技术. 他并且指出: 分析化学家作为问题的解决者, 他们有时只需确定自己所研究问题总的趋势而不需寻找信息的准确结构, 在这种时候, 对应因子分析常常可以被当作一种有效的计量化学工具来加以考虑。

俞汝勤等曾将对应因子分析应用于中国茶叶样品化学分析数据的直接处理, 以便深入研究茶叶品质与化学组成、样品来源及其他特性相互关系. 所处理的数据涉及中国 3 大茶类 (绿、红和青茶) 中 6 个品种的多个等级 (珍眉 1—7 级) 、熙春 (1—5 级) 、祁红 (1—4 级) 、凤庆滇红 (1—7 级) 、铁观音 (1—4 级) 和色种 (1—4 级). 研究结果认为, 如果设法适当增加中国茶叶中氨基酸、咖啡因和多酚的含量, 降低木质素、半纤维素的含量, 则有利于提高中国茶叶的品质.

7 因子分析的误差理论

化学工作者是完全无法获得十全十美的量测数据的，因此，自己手头所得到的带有实验误差的数据是否适合于因子分析，又如何去进行这种分析，便成了化学研究人员在接触因子分析研究之前无法回避的棘手问题．因子数目的推断（这是因子分析的最重要的一步）对于带误差的数据来说又是一个不明确、不直截了当的问题，这对于在化学研究中应用因子分析便形成了一个大障碍．

通过对数据的实验误差因子空间与数据的真实因子空间的联系的研究，现已基本上能对上述问题给出部分的答案，与此同时也已逐步形成了一套有关的误差理论．基本了解这些理论，对于加深因子分析在化学研究中的应用是很有好处的．

7.1 误差理论

如果没有实验误差存在，对于因子分析工作者来说，推演准确的因子数目是没有问题的．不过，化学工作者无法从各种量测中完全排除误差，所以，处理具有实验不确定性的数据便成为常常困扰化学研究人员的问题之一．虽然，在许多情况下，人们可以做出可信的误差估计，但对于不确定的估计却常常缺乏把握．实验误差混入因子分析流程，致使在每一个做出决定的步骤上都把过程变得复杂，鉴于此，研究实验误差是如何进入到因子分析流程中去是很重要的．只有弄清由于误差所引起的干扰，才能找到一些特殊的因子分析技术来推断因子数目以及推断实验误差．

7.1.1 实验误差的干扰

实验误差必定会产生多余的特征值和特征向量，这也必然会扰乱我们在第二章中所讨论的纯数据的因子分析过程. 对不必要的特征值的保留会复原包括实验误差在内的原始数据. 这种复原不是因子分析的初衷，也不是人们所期望的. 要排除这种干扰，首先就必须搞清楚这种干扰的来龙去脉. 由于存在实验误差，所以，每个实验量测数据点 d_{ik} 可以表示成两个项的加和

$$d_{ik} = d_{ik}^{*} + e_{ik}, \tag{7.1}$$

式中，d_{ik}^{*} 表示一个“纯”的数据点，不含实验误差，e_{ik} 是与该数据点有关的实验误差.

实验误差总会产生较纯粹空间所要求的数目要大一些的特征向量. 因此，要推断正确的特征向量的数目，就必须要有一些可供使用的可信的判据. 不过，即使是在选择了正确的特征向量数目之后，依然会发现，剔除所有的误差是不可能的. 因为，有一部分误差已混进了数据复原的流程中. 换言之，误差分成两部分

$$e_{ik} = e_{ik}^{\dagger} + e_{ik}^{0}, \tag{7.2}$$

式中，$e_{ik}^{\dagger}$ 表示混进因子分析流程中去的那部分误差，e_{ik}^{0} 是通过删除不必要的特征向量后可被除去的那部分误差，通常被叫做残余误差. 从式 (7.1) 和 (7.2)，可得

$$d_{ik} = d_{ik}^{*} + e_{ik}^{\dagger} + e_{ik}^{0}, \tag{7.3}$$

令

$$d_{ik}^{\dagger} = d_{ik}^{*} + e_{ik}^{\dagger}, \tag{7.4}$$

得

$$d_{ik} = d_{ik}^{\dagger} + e_{ik}^{0}. \tag{7.5}$$

式中，$d_{ik}^{\ddagger}$ 是采用正确数目的特征向量后的一个复原数据点．从式 (7.5) 可以看出，残余误差 e_{ik}^{0} 简单地就是原始数据点和复原了的数据点之间的差值．用于推断正确特征向量数目的判据应该以搞清楚实验误差是如何进入并渗混到因子分析流程中去的这一点为依据．

原始数据矩阵 $[D]$ 可简单地表示成两个矩阵的加和，一个是纯数据矩阵 $[D^*]$，另一个是误差矩阵 $[E]$

$$[D] = [D^*] + [E]. \tag{7.6}$$

这些矩阵的大小相同，假设为 $r \times c$ 阵，包括 r 行和 c 列（为以后讨论方便，假设 $c < r$）．然而它们的维数都不相同，纯数据矩阵 $[D^*]$ 是 n 维的，而原数据矩阵 $[D]$ 和误差矩阵 $[E]$ 的维数都为 c．

由于纯数据的因子空间是 n 维的，所以，可表示一个纯数据点 d_{ik}^{*} 成 n 个积项的线性加和

$$d_{ik}^{*} = \sum_{j=1}^{n} r_{ij}^{*} c_{jk}^{*}. \tag{7.7}$$

虽然，误差矩阵和数据矩阵大小相同，但决定误差空间比决定纯数据空间需要更大数目的特征向量，这是由于误差矩阵是由随机值构成的，只要有足够数目，任何正交的一组轴都可被选用来定义误差空间．在讨论中，用来定义误差空间的轴的数目将准确等于数据矩阵中列的数目．

可以用被称为主要轴的一套基本的轴（包含有 n 个轴）来在误差范围内描述原始数据空间．然而，要旋转误差空间，则必须要用全部 c 个轴．换句话说，还存在着一套被称为误差轴或次要轴的轴（包含有 $c-n$ 个轴），它仅与误差的残余部分有关．因为，用以描述原始数据空间的 n 个轴同时也可被用来描述误差空间的一部分，因此，与原始数据点有关的误差 e_{ik} 可被表示成如下的线性加和

$$e_{ik} = \sum_{j=1}^{n} \sigma_{ij}^{\ddagger} c_{jk} + \sum_{j=n+1}^{c} \sigma_{ij}^{0} c_{jk}, \tag{7.8}$$

这里，c_{jk} 是第 j 个轴上的第 k 个组分，$\sigma_{ij}^{\ddagger}$ 是误差矩阵的第 i 个行指定在第 j 个主要轴上的投影，σ_{ij}^{0} 是在第 j 个次要轴上的对应的投影．加和包括全部 c 个特征向量轴．

将式 (7.7) 和 (7.8) 代入式 (7.1) 中，得

$$d_{ik}=\sum_{j=1}^{n}(r_{ij}^{*}c_{jk}^{*}+\sigma_{ij}^{\ddagger}c_{jk})+\sum_{j=n+1}^{c}\sigma_{ij}^{0}c_{jk}. \tag{7.9}$$

定义 r_{ij} 如下

$$r_{ij}=r_{ij}^{*}\frac{c_{jk}^{*}}{c_{jk}}+\sigma_{ij}^{\ddagger}, \tag{7.10}$$

得到

$$d_{ik}=\sum_{j=1}^{n}r_{ij}c_{jk}+\sum_{j=n+1}^{c}\sigma_{ij}^{0}c_{jk}, \tag{7.11}$$

以此为基础，用矩阵符号来表示，则完整的因子分析解可表示成

$$[D]=[R^{\#}][C]+[R^{0}][C], \tag{7.12}$$

这里，$[C]$ 由全部特征向量构成．$[R^{\#}]$ 和 $[R^{0}]$ 则为

$$[R^{\#}]=\begin{bmatrix} r_{11} & \cdots & r_{1n} & 0 & \cdots & 0 \\ r_{21} & \cdots & r_{2n} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{r1} & \cdots & r_{rn} & 0 & \cdots & 0 \end{bmatrix}, \tag{7.13}$$

$$[R^{0}]=\begin{bmatrix} 0 & \cdots & 0 & \sigma_{1,n+1}^{0} & \cdots & \sigma_{1c}^{0} \\ 0 & \cdots & 0 & \sigma_{2,n+1}^{0} & \cdots & \sigma_{2c}^{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \sigma_{r,n+1}^{0} & \cdots & \sigma_{rc}^{0} \end{bmatrix}. \tag{7.14}$$

如果 $[R^{\#}]$ 中与次要特征值有关的零元素从这个 $r\times c$ 矩阵中被淘汰掉，便可得到正常的 $r\times n$ 行因子矩阵 $[R^{\ddagger}]$．同样，删去次要特征向量，将 $[C]$ 减小成 $[C^{\ddagger}]$，也就是正常的列因子矩阵．$[R^{0}]$ 的元素只由误差成分构成，不包含有用的信息．删除这一矩阵便导致因子

式中，r_{ij} 是第 j 个轴的第 i 个组分，$\sigma_{jk}^{\ddagger}$ 和 σ_{jk}^{0} 分别为误差矩阵第 k 个列指定在第 j 个主要轴和第 j 个次要轴上的投影.

进行与用来得到式 (7.11) 的那些过程相似的推导

$$d_{ik}=\sum_{j=1}^{n}(r_{ij}^{*}c_{jk}^{*}+r_{ij}\sigma_{jk}^{\ddagger})+\sum_{j=n+1}^{c}r_{ij}\sigma_{jk}^{0}=\sum_{j=1}^{n}r_{ij}c_{jk}+\sum_{j=n+1}^{c}r_{ij}\sigma_{jk}^{0}, \tag{7.22}$$

式中，c_{jk} 被定义为

$$c_{jk}\equiv c_{jk}^{*}\frac{r_{ij}^{*}}{r_{ij}}+\sigma_{jk}^{\ddagger}. \tag{7.23}$$

上述方程式精确地表示出实验误差是如何扰乱因子分析的特征向量成分的. 如果没有误差，σ_{jk}^{0} 和 $\sigma_{jk}^{\ddagger}$ 将为零，$r_{ij}^{\ddagger}$ 将等于 r_{ij}^{*}，c_{jk} 将等于 c_{jk}^{*}，这同所期望的恰好一样.

7.1.2 主要因子和次要因子

如果数据矩阵没有误差，那么协方差矩阵将被分解成 n 个因子的加和

$$[Z]=\sum_{j=1}^{n}\lambda_{j}C_{j}C_{j}', \tag{7.24}$$

然而，由于实验误差的存在，分解将产生较大数目的特征向量. 事实上，由于协方差矩阵的大小为 $c\times c$，这里，c 是数据矩阵的列的数目，分解将会生成 c 个因子，即

$$[Z]=\sum_{j=1}^{c}\lambda_{j}C_{j}C_{j}', \tag{7.25}$$

正如式 (7.16) 所示，这个加和将被分成两组，前面的 n 项与真实因子有关，但也包含有误差的混合，第二组则由纯粹的误差组成，正是这个第二组的项应在进一步的考虑中加以忽略.

将式 (7.18) 代入式 (7.29)，则得

$$\sum_{j=1}^{r}\sum_{k=1}^{c}(d_{ik}^{2}-{d_{ik}^{\ddagger}}^{2})=\sum_{i=1}^{r}\sum_{j=n+1}^{c}(\sigma_{ij}^{0})^{2}. \tag{7.30}$$

式中，σ_{ij}^{0} 是一个与残余误差有关的行余因子. 换句话说，原始数据与复原数据之间的平方差的加和等于次要特征值的加和. 当然，这一加和与通过忽略次要特征值而被除去的误差有关. 式 (7.30) 是一个重要的方程式，它指出了在原始数据、复原数据以及次要特征值这三者中所存在的联系.

虽然纯数据矩阵和误差矩阵不是相互正交的，但有意思的是复原数据矩阵和它的有关的残余误差矩阵是相互正交的，为证明这一点，首先考虑如式 (7.12) 所表示的完整的采用全部特征向量的数据复原. 该方程式完整地复原了数据，当然也包括所有的误差. 从式 (7.12) 和 (7.15)，可得出结论

$$[D]=[D^{\ddagger}]+[E^{0}], \tag{7.31}$$

式中，

$$[E^{0}]=[R^{0}][C].$$

在此，可注意到残余误差矩阵 $[E^{0}]$ 是原始数据阵和抽象因子分析复原矩阵之间的差. 而且，从式 (7.13) 和 (7.14) 可明显地看出

$$[R^{\#}]^{T}[R^{0}]=[0], \tag{7.32}$$

应用此方程，可发现

$$\begin{aligned}[D^{\ddagger}]^{T}[E^{0}] &= \{[R^{\#}][C]\}^{T}\{[R^{0}][C]\}\\ &= [C]^{T}[R^{\#}]^{T}[R^{0}][C]\\ &= [C]^{T}[0][C]=[0]. \end{aligned}\tag{7.33}$$

由此可得出结论：抽象因子分析复原数据矩阵与它的有关的残余误差矩阵是相互正交的. 以后，我们将要应用这一重要的事实来推导出用以确定因子空间维数的误差判据.

7.1.3 数字实例

为阐述实验误差如何扰乱主要特征向量并产生次要特征向量(它们仅由误差构成)，让我们来考察一个简单的如表 7.1 中所示的 1 因子模拟数据矩阵，这一矩阵，在表中用“纯数据矩阵”来标识，它由两个相等的数据列组成. 很明显，这个矩阵是一维的，因为对第 2 列的各对应点画第一列各点的线恰好是一条完美的直线，该直线上所有的点都精确地分布在一维的线轴上. 采用协方矩阵方法对该矩阵进行因子分析，所得结果列于表 7.2 中. 每一个行余因子 r_{i1}^{*} 是从原点至该数据点的距离，特征值 (770)，便是行余因子 (即得分) 的平方和.

表 7.1 一个模拟的原始数据矩阵及其因子分析结果

纯数据矩阵 $[D^{*}]$		误差矩阵 $[E]$		原始数据矩阵 $[D]=[D^{*}]+[E]$		复原数据矩阵 (应用一个因子) $[D^{\dagger}]=[R][C]$	
1	1	0.2	0.0	1.2	1.0	1.0936	1.1052
2	2	−0.2	−0.2	1.8	1.8	1.7904	1.8095
3	3	−0.1	0.1	2.9	3.1	2.9846	3.0163
4	4	0.0	−0.1	4.0	3.9	3.9288	3.9705
5	5	−0.1	0.0	4.9	5.0	4.9240	4.9763
6	6	0.2	−0.2	6.2	5.8	5.9671	6.0305
7	7	0.2	−0.1	7.2	6.9	7.0118	7.0862
8	8	−0.2	0.1	7.8	8.1	7.9086	7.9926
9	9	−0.2	0.1	8.8	9.1	8.9033	8.9978
10	10	−0.1	0.2	9.9	10.2	9.9974	10.1036

为弄清楚误差是如何扰乱这些结果的，采用以下的步骤. 首先，随意生成表 7.1 中所列的“误差矩阵”，然后将它加入到纯数据矩阵中以生成表中的“原始数据矩阵”，这一原始数据矩阵模拟具有实验

误差的真实化学数据. 对它进行因子分析，得到的并不是 1 个而是 2 个特征值以及有关的特征向量，这些结果列于表 7.2 中.

表 7.2　因子分析所得的特征值和行余因子

从纯数据矩阵	从原始数据矩阵	
λ_1=770	$\lambda_1^{\ddagger}$=767.514	λ_2°=0.2886124
r_{i1}^{*}	r_{i1}	σ_{i2}^{0}
1.41421	1.55487	0.14964
2.82843	2.54555	0.01344
4.24264	4.24333	-0.11901
5.65685	5.58569	0.10021
7.07107	7.00063	-0.03374
8.48528	8.48367	0.32765
9.89950	9.96895	0.26479
11.31371	11.24396	-0.15275
12.72792	12.65816	-0.14528
14.14214	14.21377	-0.13707

生成两个特征向量，那是因为原始数据点并不象纯数据点那样分布在一根一维线上，而是分布在一个二维平面上，图 7.1 对此做了说明. 在图中，原始数据矩阵的第一列的点被对第二列的相应的点作图. 这样，主要轴的方向已从它原来的 45 ° 角稍有偏移，因此 $C_1 \neq C_1^{*}$. 之所以产生这种现象，那是因为通过删除 $c-n$ 个次要特征向量也不能除去某部分的误差. 每一个行余因子 r_{ij} 是沿着主要轴至某一点的距离，数据点的垂直投影在该点同轴相交. 请注意，

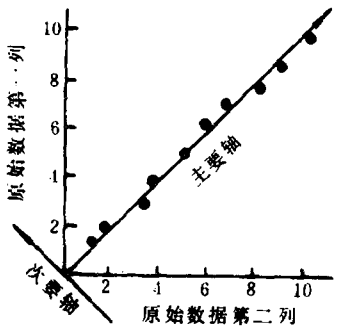


图 7.1　原始数据点与对原始数据矩阵做主因子分析所得的主要轴和次要轴之间的几何联系

原始数据矩阵的主要行余因子与纯数据矩阵的不相同，通过一些繁琐的计算便可证明：这些数据服从式 (7.10) .

所出现的第二个特征向量是由纯误差构成的，每一个行余因子 σ_{i2}^0 是沿着次要轴从原点到某一点的距离，对应的原始数据点的投影与次要轴在该点相交．第二个特征值 λ_2^0 没有物理意义，因为它是次要行余因子的平方和，而这些行余因子除了含有误差之外别无其他．删去这个次要轴，可得到复原的原始数据，它较原来的原始数据矩阵更接近纯数据矩阵．将复原数据矩阵、原始数据矩阵以及纯数据矩阵 3 者进行比较便可证明这一点．仅用主要特征向量所得的复原矩阵列在表 7.1 的最右边．

7.2 误差与分析数据的改善

抽象因子分析短路复原具有内在的统计特性，由于次要抽象特征值由纯误差构成，故消除它们自然会导致数据的改善．有意思的是，不需对起控制作用的因子有任何预备知识便可做到这一点，虽然，这并不是因子分析的初衷，但却提供了一个有价值的、意料之外的好处．在这一节中，将试图定量确定采用抽象因子分析能达到的数据改善的程度．

事实上，实验的原始矩阵是两个矩阵的加和：一个没有误差的纯数据矩阵和一个误差矩阵．在 7.1.2 节中的讨论已指出：如果对纯数据矩阵做因子分析，便可准确地得到 n 个特征值和特征向量．另方面，如如果对原始数据矩阵做因子分析，就会得到 c 个特征值和特征向量．不过，这当中，只有 n 个与真实因子有关，其余 $c-n$ 个特征值和特征向量由纯误差构成．

象对原始数据矩阵做因子分析一样，先来考察一下如果通过协方差矩阵方式对误差矩阵做因子分析究竟会出现什么情况，曾用以描述原始数据的同样的特征向量可用来描述误差空间，当然，结果生成的特征值将由纯误差组成．参照式 (7.27)，将得到下面的相类似的方程式

$$\sum_{j=1}^{c}\lambda_{je}=\sum_{j=1}^{n}\lambda_{je}^{\ddagger}+\sum_{j=n+1}^{c}\lambda_{je}^{0}, \tag{7.34}$$

式中，

$$\sum_{j=1}^{n}\lambda_{je}^{\ddagger}=\sum_{i=1}^{r}\sum_{j=1}^{n}(\sigma_{ij}^{\ddagger})^2, \tag{7.35}$$

$$\sum_{j=n+1}^{c}\lambda_{je}^{0}=\sum_{i=1}^{r}\sum_{j=n+1}^{c}(\sigma_{ij}^{0})^2. \tag{7.36}$$

在式 (7.34) 中，很明显左边项等于数据矩阵中全部误差点的平方和．因为从误差矩阵所构成的协方差阵的迹是不变的（这一点是以在分解对角化过程中所涉及的相似变换为依据的）．将这一事实同式 (7.34)，(7.35) 和 (7.36) 相结合，定义

$$\sum_{i=1}^{r}\sum_{j=1}^{c}e_{ij}^{2}=\sum_{i=1}^{r}\sum_{j=1}^{n}(\sigma_{ij}^{\ddagger})^2+\sum_{i=1}^{r}\sum_{j=n+1}^{c}(\sigma_{ij}^{0})^2. \tag{7.37}$$

上式中左边的项等于实验误差平方的加和，它也代表了误差点在全部 c 个数据列轴上的投影的平方的加和．式右边的第一个加和关系到误差点在 n 个主要特征向量轴上的投影，它代表了混进因子分析过程且不能除去的误差．右边第二项加和关系到在 $c-n$ 个次要轴上的投影，这些轴将从分析中被除去，因为与它们有关的特征值除纯误差外并无其他意义．这 3 个项均被用下面的式子来与残余标准偏差联系在一起

$$rc(\mathrm{RSD})^2=\sum_{i=1}^{r}\sum_{k=1}^{c}e_{ik}^{2}, \tag{7.38}$$

$$rn(\mathrm{RSD})^2=\sum_{i=1}^{r}\sum_{j=1}^{n}(\sigma_{ij}^{\ddagger})^2, \tag{7.39}$$

$$r(c-n)(\mathrm{RSD})^2=\sum_{i=1}^{r}\sum_{j=n+1}^{c}(\sigma_{ij}^{0})^2. \tag{7.40}$$

每一个上述表达式表示获得残余标准偏差的一种不同的方式.

将式 (7.38)，(7.39) 和 (7.40) 代入式 (7.37) 中并除以 rc，可以得出

$$(\mathrm{RSD})^2 = \frac{n}{c}(\mathrm{RSD})^2 + \frac{c-n}{c}(\mathrm{RSD})^2, \tag{7.41}$$

这一重要的等式概括了存在的理论争论. RSD 可解释为由两个项 (即，嵌入误差 (IE) 和提出误差 (XE)) 组成. 换言之，残余标准偏差 (即真实误差 RE) 可用勾股定理形式来表示如下

$$(\mathrm{RE})^2 = (\mathrm{IE})^2 + (\mathrm{XE})^2, \tag{7.42}$$

式中,

$$\mathrm{RE} = \mathrm{RSD}, \tag{7.43}$$

$$\mathrm{IE} = \mathrm{RSD}\sqrt{n/c}, \tag{7.44}$$

$$\mathrm{XE} = \mathrm{RSD}\sqrt{(c-n)/c}. \tag{7.45}$$

嵌入误差的产生是由于仅有部分来自数据的误差混入因子分析复原过程，这种误差以嵌入的状态进入到因子中去，无法通过重复的因子分析而将其除去，当次要特征向量从流程中被删去时，某些误差可被提取出来，这就是提出误差.

嵌入误差是纯数据与因子分析复原数据之间的差值的量度，而真实误差 (RE) 则是纯数据与原始数据之间的差值的量度. 为帮助记忆，我们用直角三角形来表示这些阐述，如图 7.2 所示.

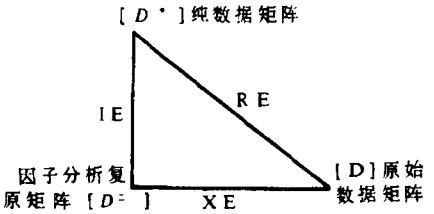


图 7.2 理论误差之间的勾股关系助记忆图

方程式 (7.44) 表明，如 $n < c$，则 IE<RE，即因子分析复原数据与

纯数据之间的误差小于原始数据和纯数据之间的本来的误差. 这样一来，采用因子分析的主要抽象特征向量，在数据矩阵中保持大于 n 的列数，我们总是能简单地使数据获得改善的 (甚至在我们不能辨别真实的基础因子时也是如此).

E.R. Malinowski 曾报导过采用合成数据矩阵进行的误差研究结果，他的发现列于表 7.3 中. 第一栏给出 n，是用来生成纯数据矩阵的因子数目，第二栏给出在每一数据矩阵中的行和列的数目，第四栏给出了被加至纯数据阵中以产生原始数据阵的误差矩阵中的误差范围. 各误差矩阵的“误差的真实均方根” (即原始数据和纯数据之间的差) 被列在第五栏中.

表 7.3　因子分析模型数据结果一览表

模拟数据和模拟误差					原始数据的因子分析		
n	$r \times c$	纯数据范围	误差范围	误差的真实 RMS	RE(RSD)	IE	$(d_{ik}^{\dagger} - d_{ik}^{*})$ 的 RMS
1	10×2	1~10	-0.2~0.2	0.148	0.170	0.120	0.087
2	16×5	2~170	-0.08~0.08	0.041	0.041	0.026	0.026
2	16×5	2~170	-0.99~0.91	0.566	0.511	0.323	0.381
2	15×5	0~27	-1.9~2.0	1.040	0.962	0.608	0.726
3	10×6	0~32	-0.9~0.9	0.412	0.376	0.266	0.311
4	16×9	-581~955	-1.0~1.0	0.548	0.499	0.333	0.398
5	10×9	-137~180	-1.0~1.0	0.463	0.372	0.277	0.400

如果对这些合成的原始数据矩阵进行抽象因子分析，采用合适数目的主要特征值，就可得到列在表中的 RE 和 IE 的值，根据因子分析所得的抽象特征向量，用式 (7.47) 和式 (7.44) 可计算 RE 和 IE，理论上，计算得到的 RE 应该等于误差的真实均方根. 比较表中第五栏和第六栏中的对应值，即可看清楚这一点是真实的 (比较时，需明白只有这些误差的第一位数才是有意义的).

表的最后一栏列出了因子分析复原数据和纯数据之间的差的均方根，这些值应该等于由抽象特征值计算所得的 IE 值，从表中可以看到，在只考虑第一位数时，情况的确基本如此. 从统计学角度来讲，在模型分析中所用的数据点的数目不大，因此，也就不能期望

得到完美的吻合.

在复原过程中采用过多数目的特征向量而使 RMS 误差成为最小，这在实践中是应该避免的，因为这会造成减小提出误差，过多的因子只会复原出不需要的实验误差. 反之，不足数目的特征向量就不足以说明起作用的真实变量，RMS 误差将会太大. 所以，必须寻找各种判据以便推断出因子的准确数目.

7.3 确定因子数目的方法

由于实验误差的存在，推断空间的准确大小是一个困难的问题，已发展用来解决这一问题的各种技术可归纳成两大类：①依赖实验误差知识的方法；②经验方法和统计学方法. 很明显，当误差情况已知时，第一类是可取的. 由于实验误差方面的信息往往缺乏，第二类方法常常是必须采用的，尽管它们具有较多的含糊性，但毕竟还是能解决问题的.

7.3.1 依赖实验误差的方法

当实验误差已知时，已研究出各种用以确定因子空间大小的判据，为便于将一种判据同另一种判据做比较，将这些判据应用于选自化学研究文献中的同一实验数据矩阵的因子分析. 在此，挑选一个由 342 个数据点构成的吸光度数据矩阵作为阐述实例，这些数据点表示 38 个 $[(en)_2Co(OH)_2Co(en)_2]^{4+}$ 溶液在 9 个不同波长处的吸光度值，吸光度值的标准偏差估计为在 0.0005 — 0.0015 吸光单位之间变动.

根据比耳定律，这一吸光度数据矩阵的秩等于吸光物质的组分数，由于可能存在水解解聚作用，这种溶液中的吸光物种数就有待于设法确定.

对该矩阵进行抽象因子分析，采用关于原点的协方差阵方式，结果得到 9 个特征向量，对应的特征值按降序排列在表 7.4 中. 在

的相关”为基础，则会得到下面的表达式

$$\mathrm{RSD}=\left[\frac{\sum_{i=1}^{r}\sum_{k=1}^{c}d_{ik}^{2}\sum_{j=n+1}^{c}\lambda_{j}^{0}}{rc(c-n)}\right]^{1/2}. \tag{7.48}$$

这些方程式为我们提供了用以推断特征值中孰属主要集孰属次要集的良好的判据. 首先，我们只以一个因子为基础来计算 RSD，作为推断过程的开始，这时，最大的，也就是最重要的特征值代表主要特征值，所有其余特征值均属于次要集，象式 (7.47) 所示那样，这些属次要集的特征值均被包括在加和中. 然后，将从式 (7.47) 计算所得到的 RSD 值同估算的实验误差相比较，如果 RSD 近似等于估算的实验误差，则表明我们已选择了合适数目的因子，这时，因子空间将是一维的. 如果 RSD 大于估算的误差，则表明我们没有选择足够数目的因子，那么，就要研究由两个最大的特征值构成主要集的情况，这时我们再次用式 (7.47) 计算 RSD 并再次将之同估算的误差相比较，如果 RSD 依然大于估算误差，就得重复这一过程，如将 3 个最大的特征值构成主要集. 如此继续这一过程，每一次将下一个最大特征值从次要集转到主要集中去，再进行计算到的 RSD 近似地等于估算误差为止，这时，我们就算正的主要集和次要集的那些特征值（以及它们

做为这一方法的实例，我们研究表 7.4 过，吸光度的误差估计在 0.0005—0.0015 之间，必须存在 3 个可辨别的因子. 因为不大于 0.001 为 0.00104，对应于 n=3，如 n=2，则对应的 RSD 这就超出了可接受的范围.

由于残余标准偏差在因子分析中是一个很有用的概念，应该搞清楚它的意义. 如果误差完全是随机的，则误差点间中将具有球面对称分布，这就意味着，数据点在任何次要量轴上的投影的平方加和的平均值都应近似地相同，因此，不

多少误差轴在计算中被应用，RSD 应是相同的．实际上，误差的分布不会是完全对称的，因子分析的主组分特性实际上是寻找那些夸大这种不对称性的影响的轴，不过，如果适当数目的因子被采用，那么，从主因子分析中计算所得的 RSD 值将会近似地等于估算误差．表 7.4 所列的 RSD 值都是由主因子分析法得到的．

2. χ 平方

有人建议，当标准偏差逐点变化而在整个数据矩阵中不维持恒定时采用 χ 平方判据．这个方法考虑到误差从一个数据点到下一个数据点的变化性，它的缺点是，必须对每一个数据点有一个合理的精确的误差估计．在这种情况下，χ_n^2 被定义如下

$$\chi_n^2 = \sum_{i=1}^{r}\sum_{k=1}^{c}\{(d_{ik} - d_{ik}^{\dagger})^2/\sigma_{ik}^2\}, \tag{7.49}$$

式中，σ_{ik} 是与可测量的 d_{ik} 有关的标准偏差，$d_{ik}^{\dagger}$ 则是用 n 个最大特征值进行的因子分析所得到的相应的数据点的值，加和包括全部实验点．对于每一套特征向量，χ_n^2 将被用来同由下式积项所表示的期望值相比较

$$\chi_n^2 \text{ (期望值)} = (r-n)(c-n). \tag{7.50}$$

应用 χ 平方判据的手续是这样的：首先，应用与最大的特征值对应的特征向量来复原数据，从这些数据按式 (7.49) 来计算 χ_1^2，然后，将此值同期望值 $(r-1)(c-1)$ 进行比较，若 χ_1^2 大于期望值，则表示因子空间是一维或大于一维的．接着，应用与两个最大特征值对应的特征向量来计算 χ_2^2，若 χ_2^2 大于期望值 $(r-2)(c-2)$，则表示因子空间是二维或大于二维的．继续上述过程直至 χ_n^2 小于它的相应的期望值 $(r-n)(c-n)$．在这一交叉点，产生与其期望值最相近的 χ_n^2 的 n 值被估计是真实的 n 值．

这一过程在表 7.4 中也有说明．在该例中，吸光度中的标准偏差对每一个点被估计在 0.0005 至 0.0015 之间变化，为计算 χ_n^2，标

准偏差中的这一离散趋势被应用于式 (7.49) 中，χ^2_n 的过渡出现在 n=2(8742>252) 和 n=3(35<210) 之间，因为 35 与 980 比 8742 与 252 更接近，所以，大概推断有 3 个因子.

3. 3σ 不吻合元素数

这一方法涉及到把观察数据与复原数据之间的不吻合数目作为所用的特征向量数目的函数的研究. 若某复原数据点，它与观察值之间的偏差比标准偏差大 3 倍或更多倍时，那么就将它归类为不吻合的元素. 须注意，这里所指的标准偏差 σ 是从实验信息中估算得到的.

表 7.4 所示的实例中，共有 342(=38×9) 个数据点. 当用 2 个特征向量复原时，有 155 个复原数据点具有大于 3σ 的不吻合，而用 3 个特征向量时，就没有大于 3σ 的不吻合数据点. 由此推断，因子空间是三维的. 这一方法的弱点是对于决定究竟多少不吻合元素数可被允许的问题，存在着随意性.

有人建议，利用这一方法来平滑数据. 那就是利用因子分析复原数据点去取代与它们相对应的那些具有大于 3σ 的不吻合的数据点. 这样平滑的目的是除去某些点所带入的误差，这些点的误差只能被认为是偶然的，在这种情况下，n 的值不要选择的太小，平滑之后，对已调节过的数据做因子分析. 这就很有希望获得更可信赖的结果，当然，这个过程带有一定的冒险性，因为人们可能在实际上强逼已调节过的数据点去服从一个比真实存在的空间更小一些的因子空间，尤其对那些真正具有独特性的点更是如此.

4. 化简误差矩阵

这是一个涉及矩阵的一系列基本操作的求秩方法. 此法以由数据矩阵 $[D]$ 的估计误差为元素组成的一个友矩阵为基础. 通过一系列基本操作使数据矩阵化简成一个等价矩阵，并使这一等价矩阵的对角元素为最大值，而在主对角线之下的所有元素均为零. 在化简 $[D]$ 的过程中，误差矩阵 $[S]$ 被连续地变换成一个等价误差矩阵，通过一系列以误差传递理论为根据的基本操作可完成这一任务. 数据

矩阵的秩就等于化简后的数据矩阵的对角元素的数目，从统计学意义上来讲，它们都是非零元素．如果一个对角元素的绝对值大于化简了的误差矩阵中对应的对角元素的绝对值的 3 倍的话，则可认为它是一个非零元素．

5. 小结

除上述方法外，尚有均方根误差，平均误差，特征值中的标准偏差等方法．如感兴趣可去参阅 E.R. Malinowski 的有关专著．

迄今所述的各种用以确定数据矩阵的真实因子空间的方法都依赖于对误差的准确估算．每一方法有可能导致一个不同的结论．对此，已有人做了详细的研究 (D.L. Duewer 等， *Anal. Chem.*, **48**, 2002 (1976))，并得出结论：判定带有不确定性的数据的真实秩并不是一件微不足道的事情．在各种判定秩的判据中，没有一种明显地是最佳的或当其单独被采用时是完全满意的，综合考虑各种判据会比信赖单一个规则能提供更好的指示．

对于将模拟误差加入至化学数据矩阵中去所造成的影响，也已有人做了详尽的研究，感兴趣者请去参阅有关文献 (*Anal. Chem.*, **46**, 821(1974)，和 **49**, 846(1977))．

7.3.2 经验方法与统计学方法

由于往往难以对误差得到精确的估算，所以，更困难的问题就变成了如何能在不依赖误差估计的条件下去推断因子空间，在这一小节中，我们将探讨各种已被建议用来解决这一关键问题的方法．

1. 嵌入误差函数

嵌入误差函数可用来在不依赖误差估计的情况下确定数据矩阵的因子数目．将式 (7.47) 代入式 (7.44)，可得

$$\mathrm{IE} = \left[\frac{n \sum_{j=n+1}^{c} \lambda_j^0}{rc(c-n)} \right]^{1/2}, \tag{7.51}$$

由上式可知，嵌入误差是次要特征值、数据阵的行和列的数目以及因子数目的函数．当进行因子分析时，这一信息总是可以获得的，当 n 取值从 1 至 c 时，可将 IE 作为 n 的函数来进行计算，随着 n 的变化，通过观察 IE 函数的情况，我们常常可以推断因子的真实数目，在数据复原中，当采用的主要特征向量越来越多时，函数 IE 将变小．然而，当我们在复原中用完全部主要特征向量并开始使用次要特征向量时，IE 将增加，因为，一个次要特征值简单地就是各误差点在一个误差轴上的投影的平方加和．如果误差是均匀分布的话，则它们在每一个误差轴上的投影应近似地相同，即

$$\lambda_j^0 \approx \lambda_{j+1}^0 \approx \lambda_c^0, \tag{7.52}$$

于是，

$$\sum_{j=n+1}^{c} \lambda_j^0 \approx (c-n)\lambda_j^0. \tag{7.53}$$

将式 (7.53) 代入式 (7.51) 就得

$$\mathrm{IE} \approx n^{1/2}k \qquad (\text{当 } n > \text{真实的 } n \text{ 时}), \tag{7.54}$$

上式中，k 是一个常数

$$k = [\lambda_j^0/rc]^{1/2}. \tag{7.55}$$

只有当我们已经在复原过程中采用了超量的特征向量时，这些方程式才适用．方程式 (7.51) 指出：一旦我们开始采用比需要的真实数目多的因子，函数 IE 实际上就会增大．在实践中，IE 的平稳的增大是罕见的，因为因子分析的主成分特性夸大了误差分布中的不均匀性，因此，次要特征值将不会是准确相等的．如果误差自始至终不是相当均匀、如果误差不是真正的随机性和如果系统误差或是离散误差存在的话，IE 函数中的最小值便不会被清楚地加以定义．

从表 7.4 所列的结果可以见到，当 n 从 1 至 3 时，IE 值出现递减，当 n 为 3 至 8 时，没有出现进一步的递减，这就证明了 3 个

因子对吸光度矩阵起作用，这一结论与上面所讲的用其他方法所得的结论是一致的.

2. 因子指示函数

这是一个经验函数，在推断合适的因子数目的能力上较 IE 函数更灵敏. 该函数 (IND) 的定义为

$$\mathrm{IND} = \mathrm{RE}/(c-n)^2, \tag{7.56}$$

这个函数象 IE 函数那样都由几个相同的变量组成，这些变量是 λ_j^0，r，c 和 n. 在研究中，人们发现：①在误差是随机的并在整个数据矩阵中分布还算均匀的情况下，如果采用正确的因子数目时 IND 函数便达到最小；②对于每一个具有过多误差的数据点，可能会生成一个额外的因子. 这常常会在 IND 函数中引入第二个最小，因此，给确定真实因子空间造成困难；③对于一个随机数据矩阵，IND 函数随着 n 增加而增大.

表 7.4 所列的结果显示出，n=3 时，IND 函数达到最小，表明有 3 个因子. 这与用其他方法所得的结论是一致的.

表 7.5 列出了 22 种醚类在 18 根色谱柱上的气 - 液相色谱保留指数的因子分析结果. 在表中，我们可以见到在 IE 函数中没有出现最小，这一异常状况可能是由先前已讨论过的任何一种或四种原因的综合影响所造成的，而在 n=6 处，IND 却清楚地出现最小. 由此可见 IND 函数可能以某种方式补偿了主成分对误差分布不均匀性的夸大.

梁逸曾等对 IND 函数做过改进并用 Monte-Carlo 模拟技术对改进前后的性能加以比较.

必须注意的是，对 IND 函数尚未有充分的了解，因此，在实际工作中，如果只单凭 IND 函数去推断因子数目有时是会令人感到困惑的.

表 7.5 对 22 种醚在 18 种色谱柱上 GLC 保留指数的因子分析结果

n	RE	IE	IND	n	RE	IE	IND
1	22.28	5.25	0.07708	10	1.40	1.04	0.02185
2	7.25	2.42	0.02831	11	1.25	0.98	0.02553
3	5.30	2.16	0.02354	12	1.07	0.87	0.03975
4	4.06	1.91	0.02070	13	0.94	0.80	0.03748
5	3.24	1.71	0.01915	14	0.73	0.65	0.04586
6	2.76	1.59	0.01914	15	0.69	0.63	0.07618
7	2.42	1.51	0.01997	16	0.61	0.58	0.15261
8	2.05	1.36	0.02045	17	0.59	0.57	0.59012
9	1.71	1.21	0.02114	-			

3. 约化特征值 (REV)

这是一个由 E.R. Malinowski 新近发展的判据. 它的原理是根据从主因子分析所得到的误差值的分布来确定其真实特征值的数目, 通过检验约化特征值 REV (即某特征值被获取它本身时所采用的自由度相除所得的商值) 的相等性和不等性, 去达到推断因子空间的目的.

要对误差矩阵 $[D]$ 做主因子分析, 必须对其列协方差阵 $[Z] = [D]^T[D]$ 进行对角化. 为了推演误差特征值的分布, 复习一下第二章中所讲过的对角化过程是必要的. 主因子分析中所用的特征值分解过程连续地获得那些说明数据最大方差的特征向量和特征值, 首先是最重要的特征向量 C_1 及对应的特征值 λ_1 (由迭代过程获得) 满足方程

$$[Z]C_1 = \lambda_1 C_1, \tag{7.57}$$

第一个剩余矩阵

$$[\mathcal{R}_1] = [Z] - \lambda_1 C_1 C_1', \tag{7.58}$$

继续迭代, 获得第二个最重要的特征向量及其相应的特征值 λ_2, 它们满足下面的方程

$$[\mathcal{R}_1]C_2 = \lambda_2 C_2, \tag{7.59}$$

第二个剩余矩阵

$$[\mathcal{R}_2] = [\mathcal{R}_1] - \lambda_2 C_2 C_2'. \tag{7.60}$$

从第二个剩余矩阵可获得第三个最重要的特征向量及其相应的特征值 λ_3 . 重复上述过程直至获得全部的特征向量和特征值.

矩阵 $[C]$ 的列是由特征向量构成的，即

$$[C] = [\ C_1 \quad C_2 \quad \cdots \quad C_c\ ], \tag{7.61}$$

同上相类似，如果对角化是对行协方差阵 $[Z] = [D][D]^T$ 进行，那么就会得到特征向量矩阵

$$[R] = [\ R_1 \quad R_2 \quad \cdots \quad R_r\ ], \tag{7.62}$$

通过主因子分析，任何一个数据阵 $[D]_{r\times c}$ (设 $c < r$) 都可被分解成 3 个矩阵

$$[D] = [R][S][C]^T, \tag{7.63}$$

式中，$[R]$ 是一个 $r \times c$ 标准正交阵，$[C]$ 是一个 $c \times c$ 标准正交阵，$[S]$ 是一个以特征值的平方根为对角化元的 c×c 对角阵 (即 $S_j = \lambda_j^{1/2}, S_1 \geq S_2 \geq \cdots \geq S_c \geq 0$). 式 (7.63) 可改写成分矩阵的加和

$$[D] = [D_1] + [D_2] + \cdots + [D_j] + \cdots + [D_c], \tag{7.64}$$

式中，

$$[D_j] = R_j \lambda_j^{1/2} C_j'. \tag{7.65}$$

通过考虑在产生每一个分矩阵中所涉及到的自由度，我们可以推演出误差分布，其理由列于表 7.6 中.

表 7.6　主成分分析所得的误差特征值的分布

偏差来源	平方的加和	自由度	约化特征值 *
$[D_1] = R_1\lambda_1^{1/2}C_1'$	对角化 $[D_1]^T[D_1] =$ 对角化 $[D_1][D_1]^T = \lambda_1^0$	rc	λ^0/rc
$[D_2] = R_2\lambda_2^{1/2}C_2'$	对角化 $[D_2]^T[D_2] =$ 对角化 $[D_2][D_2]^T = \lambda_2^0$	$(r-1)(c-1)$	$\lambda_2^0/(r-1)(c-1)$
.....	.....	.....	.....
$[D_j] = R_j\lambda_j^{1/2}C_j'$	对角化 $[D_j]^T[D_j] =$ 对角化 $[D_j][D_j]^T = \lambda_j^0$	$(r-j+1)(c-j+1)$	$\lambda_j^0/(r-j+1)(c-j+1)$

* 上标 0 表示误差

表 7.6 的第一栏列出误差数据中的偏差来源，第二栏列出每一偏差的平方加和（等于 $[D_j]^T[D_j]$ 或 $[D_j][D_j]^T$ 的对角元素的加和，当然也就等于特征值 λ_j），第三栏列出确定 $[D_j]$ 时所涉及的自由度 $(r-j+1)(c-j+1)$（因为确定 R_j 时的自由度为 $(r-j+1)$，确定 C_j 时的自由度为 $(c-j+1)$），因此误差的偏差平方的平均值，称为约化特征值 (REV_j) 由下式给出

$$\mathrm{REV}_j = \frac{\lambda_j}{(r-j+1)(c-j+1)}, \tag{7.66}$$

式中，λ_j 代表第 j 个误差特征值.

假设一个数据矩阵由随机数构成，这些随机数均匀分布且平均值等于 0，它们的取值范围为 $-x$ 至 $+x$，x 是端值. 这样的矩阵的约化特征值从统计学上讲应该是相等的，因为它们来源于相同的分布，这种分布是不受因子获取过程的影响的. 误差特征值的分布同自由度成正比，这里所指的自由度是在确定它们各自的分矩阵 $[D_j]$ 时所涉及的自由度. 误差特征值的概率分布函数 $P(\lambda_j^0)$ 表示如下

$$P(\lambda_j^0) = N(r-j+1)(c-j+1), \tag{7.67}$$

式中，N 是归一化常数.

属于真实特征向量的特征值将比从式 (7.67) 所预测到的要大，因为贡献来自真实组分．任何符合适当要求的统计学检验，如 F 检验，可被用来通过检验约化特征值的相等性而将误差特征值同真实特征值加以区别．

用两个例子来说明约化特征值判据在化学研究中的应用．一个例子涉及醋酸在碳酰区的红外谱，9 个不同浓度的样品分别在 200 个不同波长点被测量．另一个例子涉及 6 种氨基酸的混合样品 (12 个) 在 50 个不同波长处的紫外吸收．它们的因子分析结果被列于表 7.7 中．从表中可以看出，对于醋酸的例子来说，当 j 从 1 至 5 时，约化特征值 (REV_j) 不断地变小，但从 j 为 5 至 9 则基本上维持稳定，这就证明了只有 4 个真实特征值和 5 个误差特征值．对于混合氨基酸样品来说，从 j =1 至 j =7，约化特征值不断地变小，但是从 j =7 至 j =12 则基本上变化不大，这就证明了只有 6 个真实特征值和 6 个误差特征值．这些结论同用其他方法所得到的结论是相一致的．

表 7.7　醋酸在碳酰区的红外谱和氨基酸的紫外谱的因子分析结果

	醋酸			氨基酸		
j	λ_j	d.f.	REV_j	$\lambda_j(\times 10^3)$	d.f.	$\mathrm{REV}_j(\times 10^3)$
1	19.1934	1800	1.1×10^{-2}	395986.7	600	660
2	0.368079	1592	2.3×10^{-4}	795.75	539	1.48
3	0.009065	1386	6.5×10^{-6}	158.37	480	0.33
4	0.004414	1182	3.7×10^{-6}	71.09	432	0.17
5	0.000294	980	3.0×10^{-7}	7.06	368	0.019
6	0.000260	780	3.3×10^{-7}	1.80	315	0.0057
7	0.000141	582	2.4×10^{-7}	0.196	264	0.0007
8	0.000132	382	3.4×10^{-7}	0.123	215	0.0007
9	0.000099	192	5.2×10^{-7}	0.057	168	0.0006
10				0.037	123	0.0003
11				0.022	80	0.0003
12				0.019	39	0.0003

对于较小的数据矩阵的分析，REV 判据也是适用的．表 7.8 列出了两个实例的结果．一个是环己烷和环己烯混合物的质谱强度的

例子，它涉及到 4 个混合样品在 20 个 m/e 位置的质谱数据．另一个是 α- 萘酚、α- 萘胺、2,7- 二羟基萘和 2,4- 二甲氧基苯甲醛的混合物的紫外吸收谱的强度，它涉及到 9 个混合样品在 13 个不同紫外波长点的吸收．从表 7.8 中可以看出，对于环己烷和环己烯混合样品来说，根据 REV_j 值来判定，应有 2 个真实特征值和 2 个误差特征值．对于含 α- 萘酚 4 组分芳香类混合样品来说，则应有 4 个真实特征值．这些判断结果同实际情况完全相符．

表 7.8　环己烷和环己烯混合物与含 α- 萘酚 4 组分芳香类混合物的分析结果

	环己烷与环己烯混合物			含 α- 萘酚 4 组分芳香类混合物		
j	λ_j	d.f.	EV_j	λ_j^*	d.f.	REV_j
1	1035.8	80	12.947	59638.39	117	509.73
2	222.7	57	3.907	1193.88	96	12.436
3	0.7	36	0.0019	159.52	77	2.0717
4	0.2	17	0.0012	11.16	60	0.1861
5				0.0185	45	0.0004
6				0.0107	32	0.0003
7				0.0059	21	0.0003
8				0.0029	12	0.0002

* 此处得到的值均为原特征值 $\times 10^4$

4. 抽象因子分析的统计学 F 检验

特征值的加和等于数据阵中数据点 d_{ik} 的平方加和

$$\sum_{i=1}^{r}\sum_{k=1}^{c} d_{ik}^2 = \sum_{j=1}^{c}\lambda_j, \tag{7.68}$$

事实上，每一个特征值代表该特征值对应的特征向量所说明的数据中的方差．按照抽象因子分析误差理论，只有前面 n 个特征值包含有用的信息并定义因子空间，其余 $c-n$ 个特征值（叫做误差特征值）不包含有用的信息，只定义零空间（设 $c<r$）．所以，删去误差特征向量及其特征值后，主要特征值 λ 的加和就等于抽象因子分

析复原所得的数据 $d_{ik}^{\ddagger}$ 的平方加和

$$\sum_{i=1}^{r}\sum_{k=1}^{c} d_{ik}^{\ddagger} = \sum_{j=1}^{n} \lambda_j, \tag{7.69}$$

因此，原始数据同复原数据之间的关系可能概括如下

$$\sum_{i=1}^{r}\sum_{k=1}^{c} d_{ik}^{2} = \sum_{i=1}^{r}\sum_{k=1}^{c} {d_{ik}^{\ddagger}}^{2} + \sum_{i=1}^{r}\sum_{k=1}^{c} {d_{ik}^{0}}^{2}, \tag{7.70}$$

上标 0 表示在抽象因子分析中被删去的特征向量所做的贡献. 类似地可得到

$$\sum_{j=1}^{c} \lambda_j = \sum_{j=1}^{n} \lambda_j + \sum_{j=n+1}^{c} \lambda_j^0, \tag{7.71}$$

问题就是要在某些给定的显著水平之内确定数目 n . 因为特征向量是相互正交的，所以，可用方差比的表达式来区别误差特征向量与真实特征向量.

Fisher 方差比是从具有正态分布的两个独立样品组合得到的两个方差的商. 因为从抽象因子分析得到的特征向量是正交的，故能满足独立的条件，通常在实际上假定数据中的残余误差 d_{ij}^0 具有正态分布. 如果这一假设成立，则由误差特征向量表达的方差也服从正态分布. 零向量的组合方差 Var(0) 便可通过误差特征值的加和去除以组合向量的数目 $(c-n)$ 而得到

$$\mathrm{Var}(0) = \sum \lambda_j^0/(c-n). \tag{7.72}$$

真实特征向量含有既来自结构也来自实验误差的贡献，所以，真实特征向量所对应的特征值从统计学上来讲就会大于误差特征值的组合方差，于是，下面的方差比可被用来检验第 n 个特征向量（与相邻的最小特征值联系的）是否属于由较小特征值组成的零向量集

$$F(1,c-n) = \mathrm{Var}(n)/\mathrm{Var}(0) = \lambda_n(c-n)/\sum_{j=n+1}^{c} \lambda_j^0, \tag{7.73}$$

因为每一个特征值仅有一个自由度，故方差比以 1 和 $(c-n)$ 自由度为基础. 式 (7.73) 设计来检验假设 H_0: $\lambda_n = \lambda_{n+1}^0 = \cdots = \lambda_c^0$，这自然是相对于另一种假设（单侧检验） H_a: $\lambda_n > \lambda_{n+1}^0 = \cdots = \lambda_c^0$ 而言的.

采用式 (7.73)，通过下述步骤便可确定数据矩阵的真实因子数目. 首先，由最小的特征值构成零向量集，通过将相邻的一个特征值的方差与零向量集方差做比较，F 检验便可用来检验这一相邻的特征值的重要性. 如果计算所得到的 F 值小于有关 F 表中所列的 F 值 (在某些挑选的显著水平上)，则这一被检验的特征值就被加入至零集中去，然后，检验另一个相邻的特征值. 重复这种检验和加入至零集的过程，直至第 n 个特征值的方差超过 F 表中的 F 值，就可得到真实向量和误差向量的分界点.

利用式 (7.67) 所示的误差特征值分布的概率函数，可以对式 (7.73) 所表示的 F 检验进行改善. 通过约化特征值的一些性质，最终可以得到改善的 F 检验表达式

$$F(1, c-n) = \frac{\sum_{j=n+1}^{c} (r-j+1)(c-j+1)}{(r-n+1)(c-n+1)} \frac{\lambda_n (c-n)}{\sum_{j=n+1}^{c} \lambda_j^0}, \tag{7.74}$$

上式是设计来检验假设

$$\mathrm{H}_0 : \bar{\lambda}_n = \bar{\lambda}_{\mathrm{pool}}^0, \tag{7.75}$$

当然，这也是相对于另一种假设（单侧检验）

$$\mathrm{H}_a : \bar{\lambda}_n > \bar{\lambda}_{\mathrm{pool}}^0, \tag{7.76}$$

而言的. 这里 $\bar{\lambda}_n$ 代表第 n 个约化特征值，$\bar{\lambda}_{\mathrm{pool}}^0$ 代表零集之中约化特征值的平均.

用一个涉及到核磁共振质子偏移的例子来说明这一判据的应用. 该例子的数据是由 14 种溶质和 9 种溶剂的测定结果所组成的，因子分析和式 (7.74) 的应用结果列于表 7.9 中.

表 7.9　9 种溶剂中 14 种溶质的质子偏移的因子分析结果和式 (7.74) 应用结果

n	λ	$\overline{\lambda}$	F	v_1	v_2	%SL	IND$\times 10^2$
1	10243900	81300.8	51986.6	1	8	0.0	3.62
2	477.09	4.5874	10.406	1	7	1.4	2.29
3	95.46	1.1364	7.960	1	6	3.0	1.60
4	11.63	0.1762	1.401	1	5	29.0	1.93
5	8.79	0.1758	1.860	1	4	24.4	2.30
6	3.94	0.1094	1.330	1	3	33.2	3.26
7	1.56	0.0650	0.631	1	2	51.0	6.79
8	1.36	0.0971	0.83	1	1	53.0	22.43
9	0.70	0.1167	-	-	-	-	

上表中，第一栏为因子的数目，第二栏为相应的特征值，第三栏为约化特征值，第四栏至第七栏为由式 (7.74) 得到 F 检验值、与 F 检验值有关的自由分子、分母的自由度以及百分比显著水平. 最后一栏为 IND 函数值.

为了推断表 7.9 所示实例中的真实数目，我们必须鉴别全套的误差特征值. 手续是这样的：首先，指派最小的特征值为零 (即误差) 集，然后，检验相邻的最小的特征值的重要性. 如检验所得的显著水平小于某些预先规定的水平，则我们抛弃零假设 (式 7.75) 并接受另一种假设 (式 7.76). 在这一例子中，n =8 时，53.0% 是一个高的重要性，我们没有理由抛弃它的零假设，所以，将第八个特征值加至误差特征值的组合中去. 然后，相对于较小的特征值的组合检验第 7 个特征值的重要性. 从表 7.9 的百分比显著水平数据来看，我们没有理由对 n =8 至 n =4 时抛弃零假设 (在 5% 或 10% 水平)，但当 n =3 时，我们抛弃零假设，因为，这时的显著水平为 3.0%，小于 5% (或 10%). 这样一来，我们得出结论：前面 3 个特征向量对数据起作用，其余 6 个特征值构成误差向量. 从表中也可见到，当 n =3 时，IND 函数值达到最小，从 n =4 开始，约化特征值 λ 的值维持稳定. 可见，这几种判据所得的结论是一致的. 上述工作是 E.R. Malinowski 完成的.

5. 特征值比 (ER)

何锡文等提出一种新的不依赖实验误差估计的判据，叫做特征值比. 其定义为

$$\mathrm{ER}_j = \lambda_j/\lambda_{j+1}, \quad j = 1, 2, \cdots, c-1. \tag{7.77}$$

用 ER 作为 j 的函数对 j 作图，在 $j = n$ 处，曲线就会出现极大点. 根据对许多化学实例的因子分析结果的 ER_j 值观察发现，该判据还是比较灵敏的. 如表 7.10 所示，可以见到，利用 ER 判据所得结论与这些实例的研究者们用其他方法所得的结论是一致的.

表 7.10　ER 判据在许多化学研究实例因子分析中的应用结果

	1*		2*		3*		4*	
j	λ_j $(\times 10^3)$	ER_j	λ_j $(\times 10^3)$	ER_j	λ_j $(\times 10^3)$	ER_j	λ_j $(\times 10^3)$	ER_j
1	261125.7	669.7	1×10^8	2.1×10^5	19193.4	52.14	2371296.02	585.38
2	389.9	3.56	477.1	6.0	368.079	40.60	4050.90	6.73
3	109.67	5.88	95.5	8.23	9.065	2.05	601.82	203.32
4	18.64	1.46	11.6	1.32	4.414	15.01	2.96	5.19
5	12.75	7.87	8.8	2.26	0.294	1.13	0.57	1.58
6	1.62	23.02	3.9	2.44	0.260	1.84	0.36	1.38
7	0.070	1.97	1.6	1.14	0.141	1.07	0.26	1.24
8	0.036	1.04	1.4	2.0	0.132	1.33	0.21	1.23
9	0.034		0.7		0.099		0.17	1.31
10	0.006	5.68					0.13	

1* ：含 6 种氨基酸的 10 个混合样品在 50 个不同波长处的紫外吸收测量 (因子数为 6)； 2* ： 14 种溶质在 9 种溶剂中的质子核磁共振谱溶剂位移测量 (因子数为 3)； 3* ： 9 种不同浓度醋酸的四氯化碳在碳酰区 200 个波长点的红外吸收测量 (因子数为 4)； 4* ：不同浓度的 Ni^{2+} 、 Cu^{2+} 和 Zn^{2+} 组成的 10 个混合样品在 25 个波长点的可见区吸收测量 (因子数为 3)

从上表归纳出：对于 ER_j 值来说，如果仅在 j=1 时出现唯一的极大值，则因子数可推理为 1 ；如除 j =1 点之外，在其他 j 值 (如 $j = n$) 也出现极大值，则因子数应推为 n .

必须指出的是，该法在提出时未见有相应的数学证明报道.

6. Exner 函数

P.H. Weiner 等建议在不能得到对实验误差的好的估算时，可应用 Exner psi(ψ) 函数. 该函数的定义如下

$$\psi = \left[\frac{\sum_{i=1}^{r} \sum_{k=1}^{c} (d_{ik} - d_{ik}^{\dagger})^2}{\sum_{i=1}^{r} \sum_{k=1}^{c} (d_{ik} - \bar{d})^2} \times \frac{rc}{(rc) - n} \right]^{1/2}, \tag{7.78}$$

这里，$\bar{d}$ 表示实验数据的总平均，ψ 函数可从零变至无穷，最好的拟合趋近于零，等于 1.0 的 ψ 值是具有物理意义的上限值，因为，这意味着人们并没有做什么比简单地猜测每一个点具有与总平均值相同的值更好的事情. O. Exner 指出，0.5 被认为是可接受的最大的 ψ 值，因为这意味着拟合比猜测每一个点为总平均值要好上两倍. 根据 Exner 所提出的理由，ψ=0.3 被认为是一般的相关，ψ=0.2 是好的相关，ψ=0.1 则是很好的相关. 这一方法仅被指望用来对因子空间给出一个粗略的估计.

7. 交互校验法

在数据分析的某些领域中，交互校验已显示出其优越的应用性能，如在样条函数逼近中它可用来确定平滑因子 S，这类问题同因子分析问题有相似之处. 在需进行因子分析的量测数据矩阵中，包含有非随机部分，即“信号”，和随机部分，即“噪音”，后者是由“模型误差”和“量测误差”组成的. 问题就归结到要对某一因子进行初始的估算. 该因子与在一个数据集中到底有多少是信号，有多少是噪音这样的估算相对应.

在介绍交互校验法的基本思想之前，必须首先定义一个衡量模型与数据之间拟合程度好坏的拟合优度判据 (CGF). 鉴于数据分析的好的预测性质的重要性，此处应用最小二乘准则，即所有元素的一个较小的偏差平方和要比一个大的偏差平方和对应于一个较好的拟合. 交互校验的基本思想是将数据集 [Y] 随机地分成 G 个组，然后，对于一给定的初始值 $S(S = S_0)$，将第一组数据删去并得到一

个简化矩阵 $[Y^*]$，再根据矩阵 $[Y^*]$ 和选定的拟合优度准则估算出模型中的参数. 应用这些参数值及 $S = S_0$ 时的模型估算出被删除组中各样本的预测值，再由被删除样本的实际观察值与其预测值求出预测残差的平方和 (Predictive Residual Error Sum of Squares, 常简称 PRESS)，末了，对 $[Y^*]$ 加上被删除组，恢复数据集 $[Y]$. 接着，删去数据集的第二组，得到新的简化数据矩阵并对其重复上述过程，求出第二个预测残差的平方和. 类似地，删去第三组，……直到每个样本都被删除一次且其预测残差平方和均被估算过为止. 将各个预测残差平方和累加，给出一个 $S = S_0$ 时的总的“预测拟合优度判据”，记为 $D(S_0)$. 然后，S 的值从 S_0 变成 $S_1(S = S_1)$ 重复上述整个分析过程，给出 $D(S_1)$. 有次序地改变 S 的值，直至取得的 $D(S)$ 的最小值时为止. 如此时 $S = S_n$，则这个 S_n 值便被认为是给定数据集的最佳选择，也就得出一个 $S = S_n$ 的模型，它对该给定数据集具有最佳的预测性质. 至于 G 的选择，从理论上讲，有人建议使 G 等于样本容量. 但在实际上，这往往会浪费计算机时，实践证明，G 完全可以取小于样本容量的值.

关于交互校验的更详细的论述请参阅 S. Wold 等人的工作.

在因子分析技术中，有两种具体方法体现了上述基本思想. 一种叫做完全交互校验，另一种叫做二元交互校验. 前者适用于较小的数据矩阵，后者则适用于大的数据矩阵. 完全交互校验包括这样几个步骤：①删去数据矩阵的一行，得到一个简化矩阵；②对该简化矩阵进行特征分析；③采用不同维数的抽象因子空间对被删除的行进行目标检验；④将目标点与抽象因子分析预测点之间的差值进行列表；⑤重复上述过程直至每一个行都已被删除过为止；⑥对于每一行，在目标拟合中的差值均被加至表中；⑦对于不同维数的因子空间，计算差值的标准偏差 (SD).

随着真实的有意义的因子数目的增加，标准偏差 (SD) 应该减小. 当使用过多数目的因子时，SD 将不会出现更进一步的减小. 对于大的矩阵，完全交互校验法显然比较费时. 二元交互校验则可

以解决这一问题. 二元交互校验包括这样几个步骤：①先将数据矩阵分成两个子阵，其中，一个由原矩阵的奇数行组成，另一个则由原矩阵的偶数行组成；②对一个子矩阵进行抽象因子分析，然后在不同维的因子空间对另一个子矩阵的每一个行进行目标检验；③交换两个子矩阵，重复②的操作；④对于上述两次特征分析，合并目标检验所得到的差值；⑤根据上述合并，计算在不同维的因子空间时的标准偏差 (SD). SD 不再减小的点显出有意义的因子数目.

8. 百分比方差平方根变化判据 (VPVRS)

S. Alex 和 R. Savoie 根据归一化后的特征值平方根的变化情况提出了一种误差判据，由于他们没有给这种判据定义名称，在此，我们暂且把它叫做 VPVRS 误差判据. 这种误差判据的构造思想是把最后一个有意义的特征值与第一个零特征值（即没有意义的特征值）之间的空隙人为地予以加深. 对於那些在应用 E.R. Malinowski 的 IND 函数时因出现平缓的最小而损害分析过程的问题，这种思想可使方法具有更好的选择性. 另外， VPVRS 判据使用容易，不需进行复杂的数学操作. VPVRS 判据可用下面的数学公式来描述

$$V_i = (\lambda_i / \sum_{i=1}^{c} (\lambda_i)) \times 100\%, \tag{7.79}$$

式中 V_i 表示第 i 个特征值的百分比方差， λ_i 表示第 i 个特征值， c 是特征值的总的数目，然后，再根据求得的 V_i 值按下式计算 VPVRS 判据

$$\mathrm{VPVRS}(i) = (V_i - V_{i+1})/(V_{i+1} - V_{i+2}), \tag{7.80}$$

从上式可以看出， i 的取值只能从 1 至 $(c-2)$. 当 i 等于所分析的数据矩阵的秩 n 时，这一伪函数出现最大.

为了解在各种可能出现的情况中，对整个分析的可靠性程度有一个反映，他们还同时引入条件数作为一个准则. 对于一个方阵 $[A]$，条件数 $\mathrm{COND}([A])$ 可表达为

$$\mathrm{COND}([A]) = \max \lambda_i([A]) / \min \lambda_i([A]) = \lambda_1/\lambda_i, \tag{7.81}$$

式中，$\max\lambda_i([A])$ 和 $\min\lambda_i([A])$ 分别代表矩阵 $[A]$ 的最大的和最小的非零特征值. 一般地说，如果 COND($[A]$) 值较小，则该矩阵可被认为是有条件的，也就是说该矩阵没有内在的误差，故可放心地用来进行计算.

对于矩形矩阵 $[I]$，则用下式来描述其条件数

$$\mathrm{COND}([I]) = (\mathrm{COND}([A]))^{1/2}, \tag{7.82}$$

式中，$[A] = [I]^T[I]$.

从式 (7.81) 可以看出，只需要特征值与 0 不同，则 COND($[A]$) 的值就是最小的. 一旦特征值逼近于 0 或是处于因子空间之外，则 COND($[A]$) 的值就急剧增大.

S. Alex 和 R. Savoie 根据他们的研究结果认为：如果 COND($[A]$) 的值落在 1 至 250 之间，则在因子分析中易于进行成功的秩确定. 在上述区间之外时，应小心考虑所得结果. 这时，分析可能仍然是成功的，但秩的确定的可靠性却要稍差一些.

9. *其它判据*

R.I. Shrager 等利用从光谱数据得到的特征向量的相关函数来鉴别出那些高噪声的向量，噪声向量被认为是无意义的. T.M. Rossi 等利用经傅立叶变换过的向量的频率分布来鉴别高噪声的向量. 噪声向量被发现具有更高的频率成分. 他们对萤光激发 - 发射矩阵演示了方法的应用. P. Geladi 等报道了一种用以计算数据矩阵的局部秩映像的技术. 他们认为该技术对于鉴别纯组分区域、非线性区域和低信噪比区域是有用的. 有人认为此技术对于探索因子分析是一种非常有用的工具. H.M. Cartwright 等提出一种用以处理无序数据矩阵的渐进因子分析的巧妙办法 (渐进因子分析通常被认为是用于有序数据矩阵的一种曲线解析技术). 他们提出用来对特征值进行排序和作图的判据，通过对图的审察可以确定有意义的特征值的数目. X.M. Tu 等指出矩阵经奇异值分解后所获得的特征向量信息对于判断矩阵的秩是有用的. 他们提出用典型相关方法来判断带有误差的实验量测数据矩阵的秩，方法直接适用于存在有两个或两

个以上大小相同、内在因子也相同的实验数据矩阵的情况，对激发 - 发射萤光光谱数据的处理已获得成功.

7.4 误差判据的其它应用

在前一节中，我们介绍了许多种误差判据. 仔细再考察其中部分判据的特有性质，我们发现，除了可用来推断因子空间之外，它们在因子分析中还有一些其他应用，了解一下这些情况，对于在化学研究中应用因子分析是会有一定益处的.

7.4.1 推断实验误差

在 7.3.2 节中我们介绍过一些不依据实验误差估计的误差判据，它们可用来帮助推断因子数目. 当然，一旦我们知道了真实因子的数目之后，我们完全可以凭借这些判据去计算数据矩阵中的真实误差而不必依赖任何对实验误差的事先了解. 从式 (7.47) 和 (7.43) 可以得到表达式

$$\mathrm{RE} = \left(\sum_{j=n+1}^{c} \lambda_j^0 / (r(c-n)) \right)^{1/2}, \tag{7.83}$$

一旦知道 n 的值，则数据的真实误差 RE 便可计算，因为 λ_j^0 可通过因子分析而获得，r 和 c 却是数据矩阵的行和列数. 这些相同的变量都可同时在 IE 和 IND 判据中被采用.

为了说明这一点，让我们返回到表 7.4 和表 7.5 所列的例子中去. 在表 7.4 所列的关于 Co(Ⅲ) 络合物及其水解产物的吸光度例子中，几种判据综合推断其因子数为 3 . 据此，则得出真实误差 RE= 0.00104 ，这同实验误差在 0.0005 — 0.0015 吸光度单位之间的估计是一致的. 在表 7.5 所列的气 - 液相色谱分析的例子中，我们曾经得出有 6 个因子存在的结论. 对应于 n =6 的 RE 值是 2.76 ，这同实

验误差被估计为不大于 3 个单位的事实是非常吻合的.

7.4.2 检验可因子分析性

要真正具备可因子分析性，一个数据矩阵必须遵守积项线性加和的规则，同时还须由足够的数据列构成而且不带有过多的误差. E.R. Malinowski 已提出如何利用 IE 和 IND 函数来判断一个数据矩阵是否具有可因子分析性，他研究了一系列由随机数组成的矩阵. 表 7.11 中所列的是结果所得到的一个典型的实例. 从表中可见，随着 $n=1$ 至 $n=3$，IE 函数增大，这从理论上指出我们正在处理纯误差空间. IND 函数也随着 $n=1$ 至 $n=7$ 而增大，因为这一矩阵明显地不是一维的，最小必须出现在 $n=0$ 处. 当 IE 和 IND 函数表现出这类行为时，可认为数据大概是不可以进行因子分析的.

表 7.11 对一个由随机数（从 4 到 99 取值）所构成的 10×8 数据矩阵的因子分析结果

n	RE	IE	IND
1	27.00	9.54	0.55
2	24.66	12.33	0.69
3	21.88	13.40	0.88
4	18.51	13.09	1.16
5	14.42	11.40	1.60
6	11.02	9.54	2.75
7	6.47	6.05	6.47

通过研究 IE 和 IND 函数的行为，我们可以淘汰那些不可进行因子分析的数据矩阵，对于因子分析研究来说，这一淘汰过程是一种重要的辅助过程，应该当作例行工作来执行. 目前，尚未有其他方法可帮助完成这一重要任务.

7.5 目标检验向量中的误差

由于实验误差的存在，目标检验不是一个简单的、直截了当的过程. 在检验过程中，由于来自数据矩阵和目标向量本身的误差的

结合使检验过程变得复杂了，到底这些误差是如何结合的，在事先不了解误差的情况下能否将它们分离并估算它们的大小，能否发展合适的误差判据以便确定某个检验向量是否就是真实的因子，是否该检验向量将会导致数据矩阵的改善或变换，为了弄清这些问题，我们将在本节中系统地来进行推导和讨论.

7.5.1 理　　论

在 7.1 节中所阐述的论点构成我们在这里讨论的基础，从那里所做过的推导中，已经知道了原始数据阵中的实验数据是如何干扰抽象余因子的. 事实上，式 (7.10) 和 (7.23) 分别准确地显示了误差是如何对行余因子和列余因子发生影响的，如果数据矩阵中的误差足够的小，则比值 c^*_{jk}/c_{jk} 将逼近于 1，式 (7.10) 将简化成

$$r_{ij} = r^*_{ij} + \sigma^{\ddagger}_{ij}, \tag{7.84}$$

据此，可得出结论：抽象行余因子分析矩阵 $[R^{\ddagger}]$ 可表示成两个矩阵，纯行矩阵 $[R^*]$ 和对应的误差矩阵 $[E^{\ddagger}]$ 的加和

$$[R^{\ddagger}] = [R^*] + [E^{\ddagger}]. \tag{7.85}$$

从式 (3.14) 可想到，对于一给定的检验向量 $\overline{\overline{R}}_l$，$[R^{\ddagger}]$ 的元素同特征值一起被采用可产生一个最小二乘变换向量 T_l. 通过式 (3.2) 又可计算预测向量 $\overline{R}_l$

$$\overline{R}_l = [R^{\ddagger}]T_l. \tag{7.86}$$

目标检验向量中的表观误差 E_A 被定义为“预测”向量 $\overline{R}_l$ 与“原”检验向量 $\overline{\overline{R}}_l$ 之间的差

$$E_A = \overline{R}_l - \overline{\overline{R}}_l, \tag{7.87}$$

将式 (7.85) 代入式 (7.86) 中，然后将结果代入式 (7.87) 可得到

$$E_A = [E^{\ddagger}]T_l + [R^*]T_l - \overline{\overline{R}}_l, \tag{7.88}$$

如果检验向量代表一个真实因子，如果检验向量和数据矩阵两者都是纯粹的（即不带实验误差），则

$$\overline{\overline{R}}_l^* = \overline{R}_l^* = [R^*]T_l^*, \tag{7.89}$$

这里，$\overline{\overline{R}}_l^*$ 是纯检验向量．$\overline{R}_l^*$ 是纯的预测向量，T_l^* 是采用纯数据得到的变换向量．象这样的情况在化学中几乎是不存在的．不过，如果误差小，那么可期望

$$\overline{\overline{R}}_l^* \approx [R^*]T_l, \tag{7.90}$$

式中，T_l 是用原始数据得到的变换向量，因此

$$E_A = [E^{\ddagger}]T_l + \overline{\overline{R}}_l^* - \overline{\overline{R}}_l, \tag{7.91}$$

目标检验向量中的真实误差 E_T 被定义为纯检验向量与原检验向量之间的差

$$E_T = \overline{\overline{R}}_l^* - \overline{\overline{R}}_l. \tag{7.92}$$

预测向量中的真实误差 E_P 被定义为预测向量与纯检验向量之间的差

$$E_P = \overline{R}_l - \overline{\overline{R}}_l^*. \tag{7.93}$$

从上述各式，可以见到

$$E_A = E_P + E_T \tag{7.94}$$

和

$$E_P = [E^{\ddagger}]T_l. \tag{7.95}$$

将误差表示成均方根而不是表示成向量会更好一些，这一点不难做到，因为一个误差向量的内积与它的标准偏差有关．例如，在检验向量中表观误差的均方根被定义为

$$\mathrm{AET} = \left[\left(\sum_{i=1}^{r}(\overline{r}_i - \overline{\overline{r}}_i)^2\right)/r\right]^{1/2}, \tag{7.96}$$

式中，$\bar{r}_i$ 和 $\bar{\bar{r}}_i$ 分别为预测向量和检验向量的第 i 个元素，加和包括向量中的全部 r 个元素．根据式 (7.87)，上式中的加和等于误差向量 E_A 的点积

$$E_A^T \cdot E_A = \sum_{i=1}^{r} (\bar{r}_i - \bar{\bar{r}}_i)^2, \tag{7.97}$$

因此，从式 (7.96) 和 (7.97)，可得到

$$E_A^T \cdot E_A = r(\mathrm{AET})^2, \tag{7.98}$$

预测向量的均方根 (REP) 被定义为

$$\mathrm{REP} = \left[\left(\sum_{i=1}^{r} (\bar{r}_i - \bar{\bar{r}}_i^*)^2 \right) / r \right]^{1/2}, \tag{7.99}$$

式中，$\bar{\bar{r}}_i^*$ 是纯检验向量的第 i 个元素，从式 (7.93) 可见

$$E_P^T \cdot E_P = \sum (\bar{r}_i - r_i^*)^2, \tag{7.100}$$

应用式 (7.99) 和 (7.100)，可得到

$$E_P^T \cdot E_P = r(\mathrm{REP})^2, \tag{7.101}$$

同上相似地，用标准 RMS 形式，目标向量的真实误差 (RET) 被定义为

$$\mathrm{RET} = \left[\left(\sum_{i=1}^{r} (\bar{\bar{r}}_i^* - \bar{\bar{r}}_i)^2 \right) / r \right]^{1/2}, \tag{7.102}$$

从式 (7.92) 可得

$$E_T^T \cdot E_T = \sum_{i=1}^{r} (\bar{\bar{r}}_i^* - \bar{\bar{r}}_i)^2, \tag{7.103}$$

结合式 (7.102) 和 (7.103)，可得到

$$E_T^T \cdot E_T = r(\mathrm{RET})^2, \tag{7.104}$$

应用上面的推导，可将表示误差向量之间关系的式 (7.94) 转换成一个 RMS 联系．要做到这一点，我们将式 (7.94) 两边的内积看作

$$E_A^T \cdot E_A = E_P^T \cdot E_P + E_T^T \cdot E_T + E_P^T \cdot E_T + E_T^T \cdot E_P, \qquad (7.105)$$

由于这些向量的元素是随机误差，有正有负，上式右边的后两项同前两项（平方之和）比较起来相对要小一些，因此，作为很好的近似，可写成

$$E_A^T \cdot E_A = E_P^T \cdot E_P + E_T^T \cdot E_T, \qquad (7.106)$$

式 (7.98) 、(7.101) 和 (7.104) 代入式 (7.106)，可得出结论

$$(\mathrm{AET})^2 = (\mathrm{REP})^2 + (\mathrm{RET})^2. \qquad (7.107)$$

图 7.3 是这种勾股关系的一种助记忆表示．3 个不同向量（$\overline{R}$，$\overline{\overline{R}}^*$ 和 $\overline{\overline{R}}$ ）都涉及到一个因子，它们占据一个以 $\overline{\overline{R}}^*$ 为直角的直角三角形的三个角． RET 和 REP 为直角边，而 AET 为斜边．

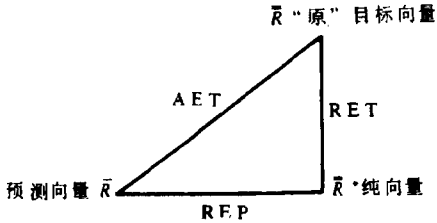


图 7.3　目标检验向量误差的勾股联系的助记忆图

目标检验向量中的表观误差往往不能直接从式 (7.96) 获得，因为检验向量可能是不完全的，仅有 p 个元素而不是 r 个元素．采用根据自由度的统计讨论，可以写出

$$\left[\frac{\sum_{i=1}^{r}(\overline{r}_i - \overline{\overline{r}}_i)^2}{r-n}\right] \cong \left[\frac{\sum_{i=1}^{p}(\overline{r}_i - \overline{\overline{r}}_i)^2}{p-n}\right], \qquad (7.108)$$

注意，n 是因子的真正数目. 将式 (7.108) 代入式 (7.96) 便可得到一个对表观误差更通用的表达式

$$\mathrm{AET} = \left[\frac{r-n}{p-n} \times \frac{\sum_{i=1}^{p} (\bar{r}_i - \bar{\bar{r}}_i)^2}{r} \right]^{1/2}, \tag{7.109}$$

这一表达式适用于目标向量完整或是不完整的情形.

为对 REP 得到一个数值，我们做如下的推导. 按式 (7.95)，E_P 的内积可用误差矩阵 $[E^{\ddagger}]$ 和变换向量 T_l 来表示

$$E_P^T E_P = \{[E^{\ddagger}]T_l\}^T \cdot \{[E^{\ddagger}]T_l\} = T_l^T [E^{\ddagger}]^T [E^{\ddagger}] T_l, \tag{7.110}$$

由于 $[E^{\ddagger}]$ 是一个 $r \times n$ 矩阵，所以，用它的转置左乘它自身将生成一个其迹等于 $\sum_{i=1}^{r} \sum_{j=1}^{n} (\sigma_{ij}^{\ddagger})^2$ 的 $n \times n$ 矩阵，根据式 (7.39)

$$\sum_{i=1}^{r} \sum_{j=1}^{n} (\sigma_{ij}^{\ddagger})^2 = rn(\mathrm{RSD})^2, \tag{7.111}$$

因此，误差的积矩阵的迹就是 $rn(RE)^2$. 与对角元素相比，这个 $n \times n$ 矩阵中的非对角元素应该小至可以忽略，因为 $[E^{\ddagger}]$ 的元素是零周围的离散误差点. 这里对角上存在 n 个元素，所以，对角元素的平均值是 $r(\mathrm{RE})^2$，因此，作为一个合理的近似，可写成

$$E_P^T \cdot E_P = r(RE)^2 (T_l \cdot T_l), \tag{7.112}$$

式中，$(T_l \cdot T_l)$ 是变换向量的点积，将式 (7.112) 代入式 (7.101)，得到

$$\mathrm{REP} = (\mathrm{RE})(T_l \cdot T_l)^{1/2}. \tag{7.113}$$

式 (7.113) 提供了一种计算预测向量中真实误差的简易方法，RE 的数值可在抽象复原步骤中从特征值计算得到，当然，该步骤也产生出因子空间的大小. 当一个检验向量被进行目标变换时，我们可对该

标向量更精确，这样的目标检验可被当作一种独特的数据改善工具来加以利用，不过，只有当目标是一个真实因子和当正确的因子数目在变换检验中被采用时，情况才是如此．然而必须注意，这样一个目标，当其被应用在组合步骤中时，倾向于损害复原数据矩阵，因为它把另外的误差引进了复原过程．

SPOIL 函数被定义为

$$\text{SPOIL} = \text{RET}/\text{EDM} \cong \text{RET}/\text{REP}, \tag{7.116}$$

如果 SPOIL 小于 1.00，组合复原数据矩阵将被目标向量所改善，如果 SPOIL 大于 1.00，那么，数据矩阵复原将被目标向量所损害．因此，在目标检验点中的误差应该是足够的小，以致于被引进因子分析过程的误差是可以忽略的，这意味着我们应该力求构成最小可能性的 SPOIL 的目标．

SPOIL 的值也为因子分析工作者判断一个怀疑的目标的总的可靠性提供卓越的判据．由于 EDM 近似等于 REP，又由于用以计算 RET 和 REP 的方程式是不准确的，所以，只希望能推导出经验的判据．通过研究模拟数据套，E.R. Malinowski 根据目标的 SPOIL 值，将目标分为可接受、尚可接受和不可接受 3 类．当其 SPOIL 值为 0.0—3.0 时，则目标被认为可接受的；为 3.0—6.0 时，被认为是尚可接受的，虽然，这个目标可能是一个真实因子，不过，在数据矩阵的复原中应用它可能会导致对数据的损害；如果 SPOIL 大于 6.0，则该目标被视为是不可接受的了，因为它会把过多的误差引入复原的数据矩阵中．

在化学研究中，确实存在着某些类型的纯的目标向量，它们不带有实验的不确定性．单位向量检验、独特性检验、对于某些官能团存在与否的检验以及结构性的向量 (如碳的数目等), 代表了纯目标的少数例子．从检验纯目标结果所得到的预测向量总是包含有由存在于数据矩阵中的实验误差所引起的误差混合，因此，不能简单地通过将一个纯目标与其预测向量作比较而做出接受或摒弃该纯目标的决定．

如 7.5.1 节所述，如果目标检验向量是纯粹的，则 RET 就为零，因此就有

$$(\mathrm{AET})^2 = (\mathrm{REP})^2. \tag{7.117}$$

在以前的推导中，计算得到的目标变换向量 T_l 被假定等于假设的纯的变换向量 T_l^* . 这一假设导致这样的结论：预测向量中的真实误差仅来源于数据矩阵的误差 (EDM) . 不用这种假设，在方程式中引入一种另外的误差贡献 (叫做过量误差)，可以更真实的描述所研究的情况.

在式 (7.89) 和 (7.90) 中，我们原来所做的假设

$$[R^*]T_l \cong \overline{R}_*, \tag{7.118}$$

应变为

$$[R^*]T_l = \overline{R}^* + E_x, \tag{7.119}$$

式中，$[R^*]$ 是假设的纯粹的抽象行因子矩阵，E_x 是过量误差向量. 因为

$$[R^*]T_l^* = \overline{R}^*, \tag{7.120}$$

故式 (7.119) 可重写成

$$E_x = [R^*]T_l - [R^*]T_l^*, \tag{7.121}$$

采用如 7.5.1 节所做的类似推导，在整个推导过程中引入 E_x ，便可得到下面的关系式

$$(\mathrm{REP})^2 = (\mathrm{EXE})^2 + (\mathrm{EDM})^2, \tag{7.122}$$

这里 EXE 是过量误差的均方根，定义如下

$$\mathrm{EXE} = (E_x \cdot E_x/r)^{1/2}, \tag{7.123}$$

我们还可得到

$$\mathrm{EDM} = (\mathrm{RE})(T_l \cdot T_l)^{1/2}, \tag{7.124}$$

根据 (7.117) 和 (7.124)，如果目标是纯粹的（即 RET=0），则得

$$(\mathrm{AET})^2 = (\mathrm{EXE})^2 + (\mathrm{EDM})^2. \tag{7.125}$$

因为 AET 和 EDM 的值容易获得，当采用纯粹的目标时，式 (7.125) 提供了一种计算 EXE 的方法.

即使是找到一个纯粹的检验向量，对于它，有 $T_l = T_l^*$，然而，误差却依然出现在预测向量中，这是因为数据矩阵中有误差所造成的. 结合考虑式 (7.115) 和 (7.125) 可知，我们所能达到的最佳的状况是

$$(\mathrm{AET})^2 = (\mathrm{EDM})^2. \tag{7.126}$$

EDM 实质上就是目标中最小可能的表观误差，从式 (7.125) 和 (7.126) 可明显地见到：过量误差实质上是对最佳状况的偏离的一个量度.

因此，目标中的过量误差与目标中的最小的可能表观误差的比值可被用来作为判断该目标的可接受性的一个判据，为方便起见，也叫做 SPOIL

$$\mathrm{SPOIL} = \mathrm{EXE}/\mathrm{EDM}. \tag{7.127}$$

如果纯粹目标是一个真实因子，则它的 SPOIL 值应趋近于零. 作为一个经验规则，当 SPOIL 值在 0—1.5 之间时，则表明该目标向量是一个真实因子；在 1.5—3.0 之间时，认为是尚可接受的，因为存在着一个增长的不确定性；大于 3.0 的 SPOIL 值时，则表明该目标是不可接受的.

除了用来判断纯目标可接受性的 3 个区间较狭窄之外，这里所讲的 SPOIL 函数的定义同先前涉及到不纯目标时的定义是一致的.

因为 RELI 函数和 SPOIL 函数都是应用得较普遍的判据，为了帮助读者较具体地了解它们的应用，在这里摘录一些实例并加以说明. 在说明中当然要结合前已讲过的目标误差理论，一则可起着一种复习的作用，另则是加深对误差理论的理解.

一个例子是 E.R. Malinowski 关于 ^{19}F 磁共振化学位移的溶剂依赖性的研究，所用的数据矩阵由 19 个刚性、非极性溶质分子溶解在

8 种溶剂中的氟位移数据构成. 通过对该矩阵进行抽象因子分析，根据 IE 和 IND 函数判据，得出有 3 个因子的结论，并得到一个等于 0.035ppm 的 RE 值，这与已知的不确定性有非常好的一致性.

根据理论探讨，溶质的气相偏移被怀疑是 3 个基本因子之一. 为了搞清这一问题，构造了一个气相检验向量，如表 7.12 所示. 这一检验向量仅含有 12 个点，另有 7 个点被自由浮动，因为它们的气相偏移没有被测量，预测向量对全部的 19 个点都产生了气相偏移. 表的底部列出了 AET， REP 和 RET 值，它们分别通过式 (7.109), 式 (7.113) 和式 (7.107) 计算得到.

表 7.12 对 ^{19}F 气相偏移进行目标检验

溶质	检验向量 $\overline{R}$	预测向量 $\overline{R}$	差值 $(\overline{R}-\overline{R})$
CF_2Br_2	-2.27	-2.50	-0.23
$CFCl_3$	-	4.98	-
CF_2ClBr	4.64	4.52	-0.12
$CFCl_2CFCl_2$	-	71.62	-
sym-$C_6F_3Cl_3$	-	118.98	-
CF_2Cl_2	12.17	11.95	-0.22
Cis-CFClCFCl	111.91	112.14	0.23
trans-CFClCFCl	-	125.62	-
C_6F_6	170.73	170.77	0.04
CF_3CCl_3	87.04	86.97	-0.07
CF_2CCl_2	95.67	95.98	0.31
CF_3CCF_3	59.29	59.41	0.12
C_4F_8	140.85	140.84	-0.01
CF_4	-	68.70	-
$C_6H_5CF_3$	-	69.94	-
$CF_3CHClBr$	-	82.66	-
α -C_6F_{14}	86.33	86.22	-0.11
β -C_6F_{14}	130.37	130.26	-0.11
γ -C_6F_{14}	126.73	126.44	-0.29
理论误差	0.19(RET)	0.05(REP)	0.19(AET)

SPOIL 的计算值为 3.8，表明这是一个尚可接受的检验向量，这暗示着蒸气偏移中的实验误差近似于溶液偏移中误差的 4 倍，这同在蒸气偏移中误差是 0.14 ppm，而在溶液中偏移的误差是 0.035

的两边便产生预测向量的期望值 $\langle\lambda\rangle$ （以特征向量值单位表示）

$$\frac{\overline{R}^2}{\sum_{j=1}^{n} t_j^2} = \frac{\sum_{j=1}^{n} t_j^2 \lambda_j^{\ddagger}}{\sum_{j=1}^{n} t_j^2} = \langle\lambda\rangle. \tag{7.131}$$

同上相类似，用概率加和去除检验向量的内积也产生检验向量的期望值．这两个期望值之间的差异 $\langle\lambda^0\rangle$ 表示预测向量与检验向量之间的方差 $\mathrm{Var}(T)$．虽然，在检验向量中有 r 个元素，但只有 $r-n-b$ 个自由度（此处，b 为空白点数目）．因此

$$\mathrm{Var}(T) = r(\overline{R} - \overline{\overline{R}})^2 / ((r-n-b)\sum_{j=1}^{n} t_j^2) = \langle\lambda^0\rangle. \tag{7.132}$$

记住，$(\overline{R} - \overline{\overline{R}})^2 = \overline{R}^2 - \overline{\overline{R}}^2$．如果将这些变量应用于行检验向量时，可得到下面相类似的表达式

$$\mathrm{Var}(T) = c(\overline{C} - \overline{\overline{C}})^2 / ((c-n-b)\sum_{j=1}^{n} t_j^2) = \langle\lambda^0\rangle, \tag{7.133}$$

式中，$\overline{\overline{C}}$ 和 $\overline{C}$ 分别为行检验向量和其预测向量．

因为上两式中，误差方差都以特征值单位表示，所以，用式 (7.72) 所示的零向量的组合方差去除这些方差，便可得到 F 检验公式

$$F(r-n-b, c-n) = \frac{r(c-n)(\overline{R} - \overline{\overline{R}})^2}{(r-n-b)\sum_{j=n+1}^{c} \lambda_j^0 \sum_{j=1}^{n} t_j^2}, \tag{7.134}$$

$$F(c-n-b, c-n) = \frac{c(c-n)(\overline{C} - \overline{\overline{C}})^2}{(c-n-b)\sum_{j=n+1}^{c} \lambda_j^0 \sum_{j=1}^{n} t_j^2}, \tag{7.135}$$

式 (7.134) 和 (7.135) 实际上是目标中的表观误差 (AET) 与来自数据矩阵的误差 (EDM) 的比值的平方

$$F = (\mathrm{AET}/\mathrm{EDM})^2, \tag{7.136}$$

在式 (7.136) 中，我们用 $r-n-b$ 或 $c-n-b$ 个自由度来定义 AET，而在以前的推导中，只是用检验向量中的元素数目粗略近似地定义 AET .

用式 (7.66) 所示的约化特征值去取代式 (7.134) 和 (7.135) 中的特征值，便可改善这两个式子. 用式 (7.67) 中所示的分布概率去除它们，就可达到这一目的，这样一来，便得到下面经改善过的统计学 F 检验

$$F(r-n-b,c-n)$$

$$=\frac{\sum_{j=n+1}^{c}(r-j+1)(c-j+1)}{(r-n+1)(c-n+1)}\quad\frac{r(c-n)(\overline{R}-\overline{\overline{R}})^2}{(r-n-b)\sum_{j=n+1}^{c}\lambda_j^0\sum_{j=1}^{n}t_j^2},\tag{7.137}$$

$$F(c-n-b,c-n)$$

$$=\frac{\sum_{j=n+1}^{c}(r-j+1)(c-j+1)}{(r-n+1)(c-n+1)}\quad\frac{c(c-n)(\overline{C}-\overline{\overline{C}})^2}{(c-n-b)\sum_{j=n+1}^{c}\lambda_j^0\sum_{j=1}^{n}t_j^2},\tag{7.138}$$

式 (7.134), (7.137), (7.135) 和 (7.138) 分别适用于列检验向量和行检验向量.

为了说明这一统计学判据的具体应用，从文献中挑选了一些数据套，它们包括质谱、红外谱、紫外可见谱、质子磁共振谱和气 - 液色谱保留体积. 根据式 (7.137) 和 (7.138) 计算各目标的 $F(v_1,v_2)$ 值. 从 F 表中查出相应的百分比显著水平 (%SL) 数据. 另外，为做比较，还计算了各目标的 SPOIL 值. 所有这些结果都列于表 7.13 中.

现以质谱数据为例来详细说明. 该套数据涉及环己烷和环己烯的 4 个不同混合物在 20 个 m/e 位置的测量值，经抽象因子分析 F 检验、IND 函数以及交互校验等判据判定有两个组分，头 5 个目

(14×9)， 14 种溶质， 9 种溶剂，目标检验以 3 个因子为基础．单位检验目标由 14 个点构成（取自 *J. Phys. Chem.*, **74**, 4537(1970) 和 *J. Amer. Chem. Soc.*, **92**, 4193(1970)）；5.气液色谱数据 (10×5)．采用行关联，目标检验以两个因子为基础．

标取自 Lorber 的工作 (*Anal. Chem.*, **56**, 1004(1984))，这些向量都含有 20 个数据点．由于环己烷和环己烯的高的显著水平 (29.4% 和 17.2%)，关于它们的零假设不能被抛弃．所以，这两个目标是可能的组分．二环 [3.1.0] 己烷的显著水平 (6.0%) 在 5% 水平是有意义的，但在 10% 水平却是没有意义的．氟环己烷和二环丙烷的低显著水平 (1.0%) 导致零假设的被抛弃，所以，这两个目标不能被接受，它们不属于因子空间．

这一节所讨论的 F 检验公式的功能还在于它允许目标检验向量可以是不完整的，即向量中可留有空白点．表 7.13 中的最后两个质谱检验目标只含有 12 个质谱点，环己烷的高的显著性 (46.0%) 指示出了环己烷的存在，而 n- 己烷的低显著性 (0.0%) 指明了它是不存在的．

关于其他数据套目标的详细情况，感兴趣者请参阅有关的文献．

7.6 因子载荷中的误差

在实际工作中，真实因子本身可能是已知的，构成目标向量的真实行余因子的数值也可能会是准确知道的．在将这些检验向量引入到目标 - 组合手续中时，就应该考虑如何去估计由组合变换结果所得到的载荷（即列余因子 $\bar{c}_{jk}$）的可信度．在这里，介绍一种以因子分析的误差理论为基础的方法．

因子载荷中的误差的均方根（用 ELF 表示）可根据以前所讨论过的误差理论来予以推导．从以前的讨论中知道，抽象列余因子可被表达成

$$c_{jk} = c^*_{jk}(r^*_{ij}/r_{ij}) + \sigma^{\ddagger}_{jk}, \tag{7.139}$$

如果误差合理地小，即 $r_{ij}^{*}/r_{ij} \approx 1$，则上式可写成

$$c_{jk} = c_{jk}^{*} + \sigma_{jk}^{\dagger}, \tag{7.140}$$

用矩阵形式表示，上式又可写成

$$[C^{\dagger}] = [C^{*}] + [E^{\dagger}]. \tag{7.141}$$

式中，$\sigma_{jk}^{\dagger}$ 是与列因子矩阵有关的误差矩阵 $[E^{\dagger}]$ 的元素．真实矩阵 $[\overline{C}]$ 可表示成

$$[\overline{C}] = [T]^{-1}[C^{\dagger}], \tag{7.142}$$

可得到

$$[\overline{C}] = [T]^{-1}[C^{*}] + [T]^{-1}[E^{\dagger}], \tag{7.143}$$

假设 $[T]$ 含有小的或没有误差，则 $[T]^{-1}[C^{*}] = [\overline{C}^{*}]$，上式就可写成

$$[\overline{C}] = [\overline{C}^{*}] + [T]^{-1}[E^{\dagger}], \tag{7.144}$$

由于 $[\overline{C}^{*}]$ 是不带误差的真实矩阵，则可知真实载荷矩阵中的误差 $[\overline{E}_c]$ 就是

$$[\overline{E}_c] = [\overline{C}] - [\overline{C}^{*}] = [T]^{-1}[E_c^{\dagger}], \tag{7.145}$$

为了得到均方根误差，考虑积 $[E_c][E_c]^T$，根据式 (7.145) 可得出

$$[\overline{E}_c][\overline{E}_c]^T = [T]^{-1}[E_c^{\dagger}][E_c^{\dagger}]^T\{[T]^{-1}\}^T, \tag{7.146}$$

由于 $[E_c^{\dagger}]$ 由有正值和负值的随机误差组成，所以，$[E_c^{\dagger}][E_c^{\dagger}]^T$ 实质上是一个对角阵．根据抽象因子分析误差理论，该对角阵中的第 j 个对角元素就为 $\sum_{k=1}^{c}(\sigma_{jk}^{\dagger}{}^2) = c(\mathrm{RE})^2/\lambda_j^{\dagger}$，这里，RE 是数据阵中的真实误差，$\lambda_j^{\dagger}$ 是第 j 个主要特征值．这样一来

$$[E_c^{\dagger}][E_c^{\dagger}]^T = c(\mathrm{RE})^2[\lambda^{\dagger}]^{-1}, \tag{7.147}$$

式中，$[\lambda^{\dagger}]^{-1}$ 是一个由主要特征值的倒数组成的对角阵．现将式 (7. 147) 代入式 (7.146)，可得到

$$[\overline{E}_c][\overline{E}_c]^T = c(\mathrm{RE})^2[\tilde{T}][\tilde{T}]^T, \tag{7.148}$$

式中，

$$[\tilde{T}] = [T]^{-1}[\lambda^{\ddagger}]^{-1/2}, \tag{7.149}$$

由于 $[\overline{E}_c][\overline{E}_c]^T$ 的第 j 个对角元素等于 $c(\mathrm{EFL})_j^2$，这里，$(\mathrm{EFL})_j$ 是第 j 个因子载荷中的误差的均方根，可得到

$$[\mathrm{EFL}]_j = \mathrm{RE}(\tilde{T}_j \cdot \tilde{T}_j)^{1/2}, \tag{7.150}$$

式中，$\tilde{T}_j$ 是 $[\tilde{T}]$ 中的第 j 行，$\tilde{T}_j \cdot \tilde{T}_j$ 是标量积.

由于 RE 可从抽象因子分析中容易地得到，T_j 可从目标因子分析中获得，因此，式 (7.150) 为计算因子载荷中的误差提供了一个简易的方法.

为了便于比较，顺便简单地介绍一下由 P.H. Weiner 等人发展出来的另一种确定因子载荷中的误差的方法 —— 大摺刀法 . 该方法主要是对一系列（如共有 r 个）简化矩阵进行完全的组合目标因子分析，每一个简化矩阵由原来的矩阵减去一个单个的行而成. 这样，便生成了总数为 r 的载荷矩阵，每一个载荷矩阵因为数据中误差的内在性质而稍有不同，这些载荷矩阵中的对应元素（即载荷）被用来对每一个载荷得到标准偏差，如果需要的话，或是得到置信界限.

应用上述两种方法来处理 10 种有机溶质的保留体积，这些溶质在溴化四乙铵的水电解液相上进行气 - 液色谱分离. 如所期望的一样，因子分析表明有两个主要因子，目标变换被用来鉴别出两个控制因子为：①每克填充料上所覆盖的液相的面积；②每克填充料的固定相容积. 从理论上讲，有关的载荷因子应分别同两个分布常数 K_A 和 K_L 相对应. K_A 是气 - 液界面的吸收常数，K_L 则是本底液体分布常数. 为了比较，将有关的数据全部列于表 7.14 中.

从表 7.14 中可以见到，这两种方法所得到的载荷中的误差是非常相近的，这一事实使我们既相信大摺刀法也相信根据误差理论所推导得到的按式 (7.150) 计算的方法.

表 7.14　用目标因子分析得到的相对应于 K_A 和 K_L 的载荷

溶质	$K_A(\times 10^{-5}\text{cm})$		K_L	
	关联 [a]	协方差 [b]	关联 [a]	协方差 [b]
四氯化碳	7.83± 0.12	7.88	3.40± 0.79	3.06
二氯甲烷	10.56± 0.22	10.55	25.51± 1.67	25.29
氯仿	30.94± 0.25	31.01	20.36± 1.17	19.97
苯	26.06± 0.21	26.03	16.80± 1.08	16.39
甲苯	76.90± 0.58	76.86	21.44± 2.46	20.72
正己烷	5.11± 0.11	5.16	0.40± 0.62	0.13
环己烷	4.32± 0.17	4.40	0.83± 1.33	0.26
正庚烷	12.06± 0.11	12.32	4.41± 4.40	3.86
2- 甲基庚烷	27.48± 0.28	27.64	1.14± 2.43	-0.22
正辛烷	34.19± 0.10	34.19	1.23± 0.86	1.23
RMS 误差	± 0.25[c]	± 0.23[d]	± 2.00[c]	± 1.4[d]

a. 采用关于原点的关联和大摺刀法得到的数据；b. 采用本节所介绍的第一种方法所得到的数据（通过关于原点的协方差）；c. 大摺刀法得到的均方根误差；d. 直接采用式 (7.150) 所得到的均方根误差.

7.7　数据实例

在 2.6 节和 3.6 节中，我们已举例演示了抽象因子分析和目标因子分析的各个步骤，不过，在进行上述演示时，数据矩阵 $[D]$（式 (2.108)）和所设计的两个检验目标 $\overline{\overline{R}}_1$ 和 $\overline{\overline{R}}_2$ 都是不带误差的“纯”数据. 在这里，我们将讨论和演示当 $[D]$ 和 $\overline{\overline{R}}_1$，$\overline{\overline{R}}_2$ 都带有误差的情况. 设一个人为的误差矩阵 $[E]$ 的值如下

$$[E]=\begin{bmatrix} 0.3 & 0.1 & 0.1 & 0.1 & 0.9 \\ -0.6 & -0.4 & 0.3 & -0.8 & 0.4 \\ 0.6 & -0.7 & -0.5 & 0.6 & -0.4 \\ -0.1 & 0.2 & -0.7 & 0.7 & 0.3 \\ -0.8 & -0.8 & -0.8 & -0.8 & 0.2 \\ 0.3 & 0.4 & 0.5 & -0.7 & -0.8 \\ 0.4 & 0.4 & -0.5 & 0.8 & -0.3 \\ -0.4 & 0.1 & 0.3 & 0.3 & 0.3 \\ 0.4 & -0.7 & 0.9 & 0.5 & 0.2 \\ 0.3 & 0.3 & -0.9 & -0.5 & -0.4 \end{bmatrix}, \tag{7.151}$$

此误差矩阵是从 -1 至 $+1$ 之间的值随机选择而得，只取至小数后第一位，这些误差的均方根为 ± 0.533．把这些误差值，按矩阵加法加至式 (2.108) 所示的“纯”数据矩阵 $[D]$ 中，就可得到一个带有误差的原始数据矩阵，设为 $[D]_{原始}$

$$[D]_{原始}=\begin{bmatrix} 5.3 & 1.1 & 6.1 & 4.1 & 12.1 \\ 17.4 & 40.6 & 26.3 & 53.2 & 60.4 \\ 18.6 & 57.3 & 27.5 & 72.6 & 65.6 \\ 21.9 & 52.2 & 31.3 & 68.7 & 74.3 \\ 7.2 & 55.2 & 15.2 & 63.2 & 40.2 \\ 8.3 & 39.4 & 14.5 & 45.3 & 33.2 \\ 18.4 & 75.4 & 29.5 & 90.8 & 71.7 \\ 20.6 & 11.1 & 26.3 & 24.3 & 57.3 \\ 28.4 & 25.3 & 36.9 & 44.5 & 80.2 \\ 27.3 & 36.3 & 35.1 & 53.5 & 80.6 \end{bmatrix}. \tag{7.152}$$

采用 JACOBI 法对矩阵 $[D]_{原始}$ 进行特征分析并按本章所述的原理算出部份误差判据，全部列于表 7.15 中.

表 7.15　对数据 $[D]_{原始}$ 分析所得结果

n	λ	RE	IE	IND	ER	REV
1	96610.99	10.388	4.646	0.649	22.42	1932.2
2	4308.51	0.519	0.328	0.0577	1147.63	119.68
3	3.7543	0.465	0.360	0.116	1.45	0.156
4	2.5971	0.415	0.370	0.415	1.50	0.186
5	1.7261					

从表中可以清楚地见到，当 $n=2$ 时，RE 值 (0.519) 小于所选择的误差矩阵 $[E]$（式 7.151）的均方根误差 0.533，IND 函数达到一个最小值，ER 值有一折点，IE 值从 $n=2$ 至 $n=5$ 时未见有进一步的减小而 REV 值则在 $n=1$ 至 $n=2$ 时显著减小，但从 $n=3$ 起却基本维持稳定．这些都说明数据矩阵 $[D]_{原始}$ 的因子数为 2．其

相应的抽象行阵 $[R^{\ddagger}]$ 和抽象列阵 $[C^{\ddagger}]$ 也就分别为

$$[R^{\ddagger}]=\begin{bmatrix} 12.8165 & 8.0609 \\ 95.4555 & 2.8905 \\ 117.6476 & -10.8471 \\ 120.0974 & 0.7695 \\ 90.4629 & -27.4889 \\ 68.9733 & -15.0555 \\ 140.2735 & -24.5478 \\ 64.5881 & 30.6779 \\ 100.6556 & 32.9906 \\ 110.2269 & 22.4169 \end{bmatrix},$$

$$[C^{\ddagger}]=\begin{bmatrix} 0.181661 & 0.439415 & 0.262683 & 0.573921 & 0.612802 \\ 0.305078 & -0.563468 & 0.302123 & -0.413880 & 0.571714 \end{bmatrix}.$$

现假设式 (3.43) 所示的两个检验目标都含有误差值，这些值分别构成对应的误差向量 $\overline{\overline{E}}_1$ 和 $\overline{\overline{E}}_2$

$$(\overline{\overline{E}}_1)'=(\ -0.1\ \ 0.2\ \ 0\ \ 0.1\ \ -0.2\ \ -0.1\ \ 0.1\ \ 0\ \ 0.2\ \ -0.1\ ),$$

$$(\overline{\overline{E}}_2)'=(\ 0.1\ \ 0\ \ -0.2\ \ 0.2\ \ -0.1\ \ 0.1\ \ 0\ \ -0.1\ \ -0.1\ \ -0.1\ ).$$

这两个误差向量的 RMS 值分别为 0.130 和 0.118，$\overline{\overline{E}}_1$ 和 $\overline{\overline{E}}_2$ 值分别按矩阵加法加至式 (3.43) 所示的相应的 $\overline{\overline{R}}_1$ 和 $\overline{\overline{R}}_2$ 中，便得到新的含有误差的“不纯”检验向量 $\overline{\overline{R}}_{E1}$ 和 $\overline{\overline{R}}_{E2}$

$$(\overline{\overline{R}}_{E1})'=(1.9\ 5.2\ 4.0\ 6.1\ -0.2\ 0.9\ 3.1\ 8.0\ 10.2\ 8.9),$$

$$(\overline{\overline{R}}_{E2})'=(\ -0.9\ \ 3.0\ \ 5.8\ \ 4.2\ \ 7.9\ \ 5.1\ \ 9.0\ \ -3.1\ \ -2.1\ \ -0.1\ ).$$

以 $[D]_{\text{原始}}$ 为基础对 $\overline{\overline{R}}_{E1}$ 和 $\overline{\overline{R}}_{E2}$ 进行目标检验，所得的各种结果列于表 7.16 中

表 7.16 以 $[D]_{\text{原始}}$ 为基础检验 $\overline{\overline{R}}_{E1}$ 和 $\overline{\overline{R}}_{E2}$ 的结果

行指定	$\overline{\overline{R}}_{E1}$		$\overline{\overline{R}}_{E2}$	
	检验值	预测值	检验值	预测值
a	1.9	1.916	-0.9	-0.937
b	5.2	5.133	3.0	2.831
c	4.0	4.022	5.8	5.964
d	6.1	5.999	4.2	4.054
e	-0.2	0.031	7.9	7.873
f	0.9	0.967	5.1	4.989
g	3.1	2.938	9.0	9.105
k	8.0	8.066	-3.1	-3.016
i	10.2	10.200	-2.1	-2.156
j	8.9	8.978	-0.1	-0.006
最小二乘	0.04893		0.03486	
变换向量	0.15990		-0.17170	
AET	0.10475		0.10984	
REP	0.08677		0.09091	
RET	0.05868		0.06165	
目标中已知的 rms	0.130		0.118	
SPOIL	0.67632		0.67810	
RELI	100%*		100%*	

*. 因为 $(\text{RET})_{\text{est}}$ 大于 RET，故检验被认为 100% 可信.

8 因子分析在组分分析中的应用

对所研究体系的定性和定量的了解是化学家们普遍关心的问题之一. 在这方面, 因子分析是一种颇具特色的手段, 它能帮助化学家去顺利地解决许多单独用传统的分析化学手段所不能满意解决的难题. 在本章中, 我们将扼要地向读者介绍自 60 年代以来化学家们如何在各种分析手段中结合各种因子分析法去解决这方面的问题.

8.1 吸收光谱

吸收光谱是因子分析一直应用得比较普遍的研究领域之一. 由于某些种类吸收光谱的获取不需复杂的设备和昂贵的经费, 而且, 数字化后的光谱又符合因子分析的要求, 因此, 多年来, 因子分析在吸收光谱研究领域中的应用一直是比较活跃的研究课题, 这种应用反过来大大地增强了吸收光谱解决问题的能力. 在这里, 着重介绍紫外可见光谱, 红外谱和近红外谱等方面的内容.

8.1.1 紫外 - 可见光谱

最早把因子分析法用于确定混合物的吸光物种数的是 J.J. Kankare, 他在前人工作的启发下, 用因子分析法研究了络合物在水溶液中的平衡问题. 在 230 — 360nm 之间以 5nm 为间隔, 测定了 17 份由 8×10^{-5}M Bi^{3+}, 1M 高氯酸以及不同含量的氯化钠和高氯酸钠组成的溶液 (维持离子强度恒定) 的吸光度. 通过对原始的吸光度数据矩阵进行分析, 确认有 7 种吸光物质存于溶液中, 即: Bi^{3+}, $BiCl^{2+}$, $BiCl_2^+$, $BiCl_3$, $BiCl_4^-$, $BiCl_5^{2-}$, $BiCl_6^{3-}$. 他利用在主因子分析中重构的数据取代那些具有过大误差的原始数据

点，改善了测量数据．用改善了的数据重新分析，得到更可靠的结果．他还结合其他方法计算出浓度矩阵，然后通过因子分析法求出吸光系数矩阵而获得各络离子的纯组分光谱． Z.Z. Hugus 和 A.A. El-Awady 研究了钴（Ⅲ）双核配合物的吸收光谱，得出其水解解聚有 3 个组分的重要结论． E.R. Malinowski 对 Hugus 的测量数据进行分析，不仅得出有 3 种水解物质的相同结论，而且还计算出实验误差．

P.C. Gillette 等提出了一种用于分析紫外光谱数据的因子分析算法，并且用合成数据进行验证． Malinowski 和 McCue 用目标因子分析方法在 260 — 280nm 范围内测定了二甲苯混合物的光谱，对其进行研究，不仅确定了混合物体系中的物种数，而且还给出各组分的含量．

H. Gampp 等人用渐进因子分析方法处理光谱滴定数据和研究络合物的平衡常数．详细内容在第五章中已有叙述．

李通化用因子分析法结合光度分析，对转氨酶的组成进行同时测定．何锡文等人将因子分析与光度分析相结合，提出了用光度 - 因子分析法进行多组分的同时测定的方法，通过解析多种金属离子混合物与同一显色剂作用生成多组分络合物体系的紫外 - 可见光谱数据，可求出各种金属离子的含量．他们还用改进的因子分析法，同时测定药物止痛片中 4 种组分的含量，获得了满意的结果．

通过评估二次矩阵的特征向量， A. Meiter 采用主成分分析法从色素混合物吸收光谱中提取组分谱．

P.J. Gemperline 等曾用主成分回归（主成分分析与多元回归两种技术的结合）来完成定量多组分光谱分析的背景校正．所用标准和待测样品中均含有相同的背景组分．他们用硝酸镍、硝酸钴和硝酸铬 (III) 水溶液的 UV 谱试验所提出的方法．结果发现，虽然硝酸镍和硝酸铬 (III) 的 UV 谱严重重叠，但在强吸收铬背景存在时依然可获得镍和钴的准确浓度测定结果．方法还用来测定有弱吸光性和光散射的药片赋形剂存在的情况下盐酸假麻黄素和盐酸 triprolidine

的浓度.

K. Nakanishi 等量测了被稀释在 3 种类型的混合溶剂（乙醇 - 庚烷、1,4- 二噁烷 - 环己烷和四氢呋喃 - 环己烷）中的碘的电子光谱. 量测时，碘的电子光谱作为混合溶剂中各溶剂的摩尔组成的函数. 然后，用因子分析技术去处理光谱数据，结果发现在醇溶液中不同的碘 - 溶剂络合物的数目小于在醚溶液中的数目.

潘忠孝、夏四清等应用目标因子分析法成功地实现了 4 种氨基酸（酪氨酸、色氨酸、苯丙氨酸和二羟基苯丙氨酸）混合体系中的吸光物种数、物种和各物种组分含量的同时测定. 受到这种成功的鼓励，他们又把氨基酸的物种数从 4 增至 6，再加入胱氨酸和组氨酸. 同样也获得了非常满意的结果（混合体系中，这些组分的浓度范围为 0—25ppm）. 上述氨基酸的紫外吸收谱 (281.8—230nm) 严重重叠，如不借助象因子分析这样的计量化学手段，想在不做任何分离的情况下直接采用传统的分光光度法去分析它们的混合体系，那是无法实现的. 为了进一步证明因子分析对有机物混合体系同时测定的适用性，他们又完成了六种芳香类化合物（α- 萘酚、α- 萘胺、2,7- 二羟基萘，2,4- 二甲基苯甲醛，水杨酸甲酯和邻苯二甲酸二丁酯）混合体系的同时测定（各组分的浓度范围为 0—10ppm）. 这些芳香类的化合物的紫外吸收谱（193.95—306.45 nm）相互重叠也非常严重. 在这些工作中，他们发现，解决这类问题时，所选用的连续吸收波长范围对计算结果产生非常严重的影响（但对于象苯酚和间苯二酚这样的简单混合体系，虽然，两者的紫外吸收重叠也非常严重，这种影响却甚微），而取样波长间隔的改变则对计算结果无甚影响，对这种现象的解释，尚有待于今后的进一步研究.

8.1.2 红外光谱

G.T. Rusmassen 及其合作者利用主因子分析法研究了人工合成的二甲苯混合物在 500 — 3500 cm^{-1} 的傅里叶变换红外光谱，准确地确定了二甲苯混合体系中所含的物种数. M.K. Antoon 等研究

了聚合薄膜的 FT-IR 光谱，对 7 份由不同比例的无规聚苯乙烯和聚 2,6- 二甲苯 -1,4- 苯氧组成的薄膜样品的指纹区进行研究，结果表明：有 3 种物质存在，从而说明其中某一种组分聚合物在膜中发生了构象变化，并且与该组分的含量呈函数关系．对半晶态的聚乙烯对苯二甲酸酯的红外光谱分析表明，它由相应的两个组分（晶态的和非晶态的）组成．研究聚氯乙烯薄膜的热处理行为，发现有 8 种组分存在，说明是由于链中构型和构象的无序性造成的．

J.T. Bulmer 等为研究乙酸的单体一环状二聚物平衡问题，在乙酸 - 四氯化碳溶液中研究其羰基区域的红外光谱，结果发现有 4 种吸光物质存在．这一结果推翻了以往那种一直认为只存在单体乙酸和环状二聚体两种物质的结论，从而给出在该体系中还存在含有氢键物质的证据．受研究乙酸问题成功的鼓舞，他们又对三氯乙酸 - 四氯化碳体系进行研究，结果表明，和乙酸一样，体系中存在 4 种物质．这一结论既无法通过简单的谱带外形分析得到，也无法通过浓度研究单体 - 二聚物平衡常数的方法而获得．

J. Korppi-Tommola 等用因子分析法研究五氯苯酚和丙酮在四氯化碳溶液中的络合平衡，分别对羰基和羟基区域的伸缩振动的红外光谱进行研究，结果两者都指出有 3 种物质存在于平衡体系中．

谱带分辨研究往往需要对谱带的形状有某些假设，如高斯型和洛伦兹型．但实际上，这种假设往往并不正确． W.H. Lawton 等提出了一种基于因子分析的解析重叠谱带的方法．该法无需对谱带形状作假设，能把重叠的谱带分解为各自原来的形状． N. Ohta 把这一方法用于研究照相染料的成分．

当向染色剂水溶液中加入聚合电解质时，往往在较低的波长范围内出现新的吸收峰． K. Yamako 等提出了一种研究这种现象的因子分析方法． C.H. Lin 及其同事研究了一种光谱自动解析因子分析方法，成功地用于分析混合物光谱．李科等人发展了一个因子分析软件，包括主因子分析、非正交变换以及目标检验程序，并用质谱和红外光谱数据对软件作了验证．

S.R. Culler 等对 E- 玻璃纤维板表面上的硅烷偶联剂的傅立叶变换红外光谱进行因子分析以确定纯组分的数目和提取纯组分光谱、改善收集到的光谱信噪比以及测定未知混合物的浓度.

8.1.3 近红外光谱

近红外区的频率范围为 12500 cm^{-1} (800nm) 至 4000 cm^{-1} (2500nm)，在这一区域中，可发现许多产生于某些基本谱和组合谱的简谐运动的吸收带，这些基本谱和组合谱往往都同氢原子有联系. 在这些吸收带中，首先碰到的是 O-H 和 N-H 分别在 7140 cm^{-1} (1400nm) 和 6667 cm^{-1} (1500nm) 处的伸缩振动以及在 4548 cm^{-1} (2200nm) 和 3850 cm^{-1}(2600nm) 处由 C-H 伸缩和烷基变形所形成的组合谱. 水在近红外区的吸收带在 2760， 1900 和 1400 nm，在分析实践中要采用那个吸收带，这取决于被分析的物质的浓度.

如同在其它种类的吸收光谱研究中一样，因子分析和其它多元分析技术的采用将会大大地增强近红外光谱的功能且可拓宽其应用领域.

近红外光谱同主因子分析技术相结合已被 D. Bertrand 等成功地应用于预测草饲料蛋白质含量和生物体外干物质的消化率.

在上述研究中，应用主因子分析回归技术来开发一个适用于用近红外光谱测定草饲料成分的预测方程式，这一技术包括两部分内容：用主因子分析来建造新的合成变量；用这些新的变量进行多元线性回归.

从作物选择项目的实验中获得某地区的 345 个青草样品，它们的种植条件及收获日期不同. 将它们分成校正集和预测集. 样品先在炉中 80 ℃下烘干 24 小时，然后在锤磨机中粉碎. 称取一定量样品置于尼龙袋中，加入稀的反刍动物消化母液在 39 ℃保温 24 小时，然后流水洗涤，再在胃蛋白酶中保温 48 小时，然后干躁称重，这样便可测出体外消化率. 至于样品中的粗蛋白组分的测定，则采用美国官方分析化学家协会 (AOAC)1975 年所提出的方法来进行. 这些

结果可对预测结果提供比较.

研究结果证明，采用主因子分析变量得到的预测方程式同那些用传统方法获得的预测方程式一样精确. 由于主因子分析变量是线性组合的，故所得到的预测方程式可采用形式为 $\lg(1/R)$ 的光学数据，这就允许研究时通过常规的滤光分光光度计来使用所得到的软件. 在相关圆上的参数（波长和化学数据）的主因子分析表达式是一种估算近红外分光光度法预测化学数值的能力的快速方法，如果在一个样品集中要研究几种组分，则可以同时对于每一种组分进行估算. 此外，采用主因子分析变量只需要很少的计算机时. 如果在校正集中的样品数小于 100 左右，那么只要用一个不太长的计算机时便可处理非常大量的光谱数据.

近红外反射分析是一种快速和准确的方法，在工业上已应用了近 30 年了，某些组分，如水、蛋白质、淀粉和纤维等都常用这种技术进行分析. 近来，D. Bertrand 等借助主因子分析技术，采用近红外反射技术来进行小麦品种的鉴定. 小麦品种鉴定的主要困难来自小麦的近红外反射谱的相似性，与由粒度影响而产生的差异比较，不同品种的小麦的光谱差异是非常小的. 不过，在实践中，需要加以鉴定的不同小麦的品种数目是不大的，尤其是对于某一个加工厂来说，小麦的供货者或是供货品种也不是太多. 可用于小麦品种鉴定的实验手续包括以下几个步骤：①减小粒度对光谱变化的影响；②选择有意义的波长；③光谱数据的主因子分析；④多元判别分析.

将小麦样品控制在含水量为 16.5%，然后磨制样品，使出粉率在 66—72% 之间. 用 Technicon Infraanalyser 500 分光光度计从波长 1100 nm 起记录红外反射谱，测量步长为 8nm . 光谱吸光度值被转移至 IBM-PC 微机上并进行数据处理.

该手续用于 6 个不同小麦品种的鉴定. 同电泳技术比较，近红外反射法准确度要差一些，然而，它的简易和迅速可能会在工业流程的过筛工序中引起兴趣. 例行的品种鉴定适用于数目不大的品种之间的鉴别，尤其适用于剔除某些已知的不希望有的品种.

此外，在近红外谱应用研究中结合采用主因子分析技术和其它多元分析技术还可解决诸如判别面包烘烤质量 (M.F. Devaux 等)、研究甜菜根浆在反刍动物瘤胃中的微生物降解（研究残渣中的蛋白质和果胶）和预测苹果中的糖含量 (P. Robert) 等. 在理论研究方面，还可以用来从近红外谱数据中提取因子判别式光谱模型等 (M.F. Devaux).

I.A. Cowe 等曾报导过应用近红外和主组分分析技术来测定小麦面粉中的水分百分率和蛋白质百分率以及在磨碎了的大麦中的热水提取值，热水提取值被用来判断大麦的发芽质量. 此外，他们还用这一技术来测定磨碎的和完好的油菜籽中的油的浓度. 用完好油菜籽的结果较用磨碎的油菜籽的结果要好，但在预测油菜籽的蛋白质百分含量方面较少成功.

S.M. Donahue 等对甲烷、乙烷和丙烷的高压合成天然混合物的近红外反射谱应用主成分回归技术，以便检验用光谱法监视天然气热值的可能性，天然气的热值是输送管道的操作者、配气站和大工业用户最为关心的问题. 在研究中，他们比较了 4 种计算方法 —— 伴随光波长选择的逆比耳定律、主成分回归、偏差最小二乘和用傅立叶变换预处理过的光谱的主成分回归. 用预测的标准偏差来判断上述 4 种技术在不同的压力 (100， 250， 500， 750 和 1000 Psi) 下的运行情况. 他们的研究结果报道：第四种方法给出最好的结果.

G. Puchwein 通过对近红外光谱的因子分析来选择标准试样. 对吸光度进行因子分析之后，样品的得分可用来界定因子空间的区域，该区域的扩展和各数据点彼此之间的距离可用作反复地剔除不重要试样的判据. 他应用这种方法去选择测定玉米和油菜籽中的蛋白质、水份和油含量时的校准试样.

H. Mark 曾详细地讨论了多元回归、判别分析和主成分分析这 3 种算法在近红外光谱分析中应用时各自的特点和局限. 他认为不应该笼统地讲这些算法中“谁好谁坏”，而应该针对不同的问题类型适当地选用不同的方法.

综上所述，因子分析已在处理吸收光谱数据中得到相当广泛的应用．借助因子分析方法对混合物的紫外 - 可见光谱数据解析，可确定络合物体系中的络合物（或络离子）种数及其平衡常数；对多组分体系中各组分的相应含量进行多组分同时测定；消除干扰组分以及背景的影响等．

对混合物红外光谱进行因子分析还可确定同系物及同分异构体体系的组分数；研究高聚物的组成及构象转化；研究单体 - 聚合物的络合平衡；可对混合物的红外光谱解析以获得单组分的纯光谱并改善光谱的质量；研究有机物的相转移问题；更值得注意的是利用因子分析还可研究光谱 - 特性关系问题．

8.2 发射光谱

对于确定对应于一个发射光谱的组分的数目来说，因子分析法已被证明是一种独特的手段．在喇曼光谱、萤光光谱等这一类方法中，发射强度既取决于浓度同时也取决于每一发射物种的独特光谱性质，因而，因子分析法比较容易地被加以应用．

8.2.1 喇曼光谱

T. Jarv 等在对氯化铟（Ⅲ）水溶液的激光喇曼光谱的研究中，观察到一个宽而不对称的单独喇曼谱形，它具有一个随着 Cl^- 与 In（Ⅲ）的浓度比 (R) 增加而从 311 至 279 cm^{-1} 产生漂移的最大值．这就暗示着这一单独的谱带可能是几种 In—Cl 离子型体的混合物．这一问题对于因子分析来说是很理想的，因为喇曼强度 I_{ik}（第 k 份溶液在第 i 个波长处观察到的强度）符合下面与比耳定律相类似的表达式

$$I_{ik} = \sum_{j=1}^{n} J_{ij} C_{jk}, \tag{8.1}$$

式中，J_{ij} 是第 j 个型体在第 i 个波长处的克分子强度，C_{jk} 是第

j 个型体在第 k 份溶液中的浓度.

以数字格式记录 31 份不同 R 值的溶液的喇曼光谱，在波长范围为 170—410 cm^{-1} 期间对每一条谱进行扫描，进行背景校正和数字化后得到 481 个点. 对最后得到的 481×31 数据矩阵进行因子分析，结果证明有 4 个型体存在. 这一结论通过考察残余标准偏差，χ^2，大于标准偏差 4 倍的不吻合元素数等判据而被证实.

因子分析法然后又被用来估算那些少于 4 个组分存在的 R 的范围. 结果表明了当 R 值小于 2.36 时仅有两个型体存在. 这同以半半波宽度和不完全第三矩为依据的其他估算不相一致. 它们指出在这一范围内有 3 个型体存在，这一差异可能是在这一范围内 3 个物种中的两种的强度之间的偶然线性所造成的结果，这种偶然的线性使得这些物种不能为因子分析所辨别. 4 个物种 ($[InCl(H_2O)_5]^{2+}$，$[InCl_2(H_2O)_4]^{+}$，$[InCl_3(H_2O)_3]$ 和 $[InCl_4(H_2O)_2]^{-}$) 被假定与光谱相关. 鉴于有限的精度和非常严重的谱带重迭，不能估算这些物种之间的平衡常数. 这一研究指出，对喇曼光谱中的一个单独谱带的观测不足以证明一个物种的存在.

对 $ZnCl_2$ 和 HCl 的水溶液混合物的喇曼光谱进行过研究，采用不同的 Cl^- 和 Zn^{2+} 浓度比，在 Zn–Cl 延伸域记录光谱. 因子分析指出只有两种光散射组分存在，它们被假定是 $ZnCl_4^{2-}$ 和 $ZnCl_2$. 采用谱带分辨技术，反应 $ZnCl_2 + 2Cl^- = ZnCl_4^{2-}$ 的平衡常数被估算为 0.22 M^{-2}.

E.R. Malinowski 等用因子分析技术来对硫酸水溶液组分的喇曼谱进行分离. 抽象因子分析被用来确定因子数，关键集因子分析被用来鉴别每一组分所独有的光谱波数，光谱分离因子分析揭示了每一个未知组分的光谱，目标因子分析则可解决各光谱组分的相对量，根据因子载荷所得到的浓度分布以及分离光谱均可被用来鉴别化学物种.

T. Ozeki 等用因子分析去研究在酸性水溶液中钼的同多酸盐的重叠喇曼光谱和化学平衡. 发现在 pH 7.2 — 2.1 之间钼的同多酸溶

溶体系中存在 4 种型体：单体、七聚体、质子化七聚体和八聚体. 七聚体型体对应于仲钼酸盐，八聚体对应于八钼酸盐. 他们并获得了这些型体的平衡常数.

J.B. Ng 等用因子分析和傅立叶自重叠合法去处理氘化了的醋酸的喇曼谱. 研究羰基的伸缩区域，发现线性无关的光谱组分数目为 4.

8.2.2 萤光光谱

在此之前，本章中所讨论的应用都要求要有一个由不同含量的相同组分所构成的一系列混合物的数据矩阵，某些例外的情况也要求组分要存在于化学平衡中以便温度或 pH 的改变能产生组分的改变. I.M. Warner 等人巧妙地提出了一种用来在一个单个的混合物中测定萤光组分的数目及各组分独自的萤光光谱的方法. 因子分析 - 萤光技术不要求组分之间的平衡.

然而，仅当数据矩阵包含有只有某一组分才会产生吸收和发射的波长区域时，才有可能获得在一个混合物中的该单独组分的完整的萤光光谱，这种分析方法是以这样的事实为依据的：每一种萤光组分都被它自身的萤光强度对激发波长 λ_i 和观察的发射波长 λ_j 这两个参数的唯一依赖关系所表征. 数据矩阵是一个激发 - 发射矩阵 $[M]$，它的元素 M_{ij} 是在 λ_i 激发时在 λ_j 处测得的萤光强度. 对于稀的混合物，这些强度取决于与每一个萤光组分 k 有关的积函数的加和

$$M_{ij} = \sum_{k=1}^{n} \alpha_k X_{ik} Y_{kj}, \tag{8.2}$$

式中，α_k 与组分 k 的浓度成正比，X_{ik} 与每单位浓度的 k 在 λ_i 处吸收的光子的数目成正比，Y_{kj} 与组分 k 在波长 λ_j 处发射的萤光的分率成正比. 在这里需注意，X_{ik} 与 λ_j 无关，Y_{kj} 与 λ_i 无关. 上述表达式对于因子分析来说是较理想的.

I.M. Warner 等人在一系列试探性的研究中，对 10 个二组分激

光 - 发射矩阵进行因子分析. 这些矩阵涉及到 5 种芳香烃：蒽、芘、芘、䓛和萤蒽. 他们先保持激发波长不变而扫描发射光谱，然后将其在 50 个波长处数字化，再对这些波长处的强度进行格式化后直接将它们送入计算机中，然后，改变激发波长，在相同的 50 个波长处重复上述步骤，如此重复 50 次，直至获得 2500 个数据点. 每一次扫描产生数据阵的一行，这样所获得的数据形成一个 50×50 激发 - 发射矩阵.

通过记录纯溶剂的萤光发现一个散射光组分. 首先从每一个数据矩阵中扣除掉散射光贡献和估算的暗电流，然后，为了校正散射光贡献，再减去已被扣除过的矩阵的一个倍数. 对这些经预处理过的数据矩阵进行抽象因子分析，10 个不同的数据矩阵中的每一个的秩都正确无误地等于萤光组分的数目，不需借助于纯组分或是它们的光谱的任何事先了解就可推导出纯组分的萤光光谱.

测定一个萤光混合物定量组成的传统步骤涉及到对一套联立方程组拟合数据，它要求要知道存在于混合物中全部物种的种类和各自的萤光光谱， C.N. Ho 等人发展了一种秩消法（见第四章）. 该法不要求对其他萤光组分的鉴别就可得到混合物中的单个的萤光组分的定量构成. 前面的章节中已详细地介绍过秩消因子分析法，在此，结合 Ho 等人的工作概括一下应用于萤光光谱研究中的秩消法的基本原理：一个多组分混合物的激发 - 发射矩阵 $[M]$ 有一个等于 n_c（存在的组分的数目）的秩，一个纯组分的对应的激发 - 发射矩阵 $[N]$ 的秩在理论上等于 1，如果从 $[M]$ 中减去 $[N]$ 的正确的量，将得到一个简化了的矩阵 $[L]$，它的秩等于 $n_c - 1$. 要达到这一目的而必须被从 $[M]$ 中减去的 $[N]$ 的量应等于 $(c_k/c_k^0)[N]$. 在这里，c_k^0 是在相同的溶剂中被用来获得 $[N]$ 的纯组分 k 的浓度，换言之

$$[L] = [M] - (c_k/c_k^0)[N], \tag{8.3}$$

应该注意的是，即使两个矩阵 $[M]$ 和 $[N]$ 都已被对于暗电流（或无照电流）和由溶剂引起的光散射进行校正，然而，随机噪声会混淆秩约简过程. 确定正确的 c_k/c_k^0 值的有效方法是把 $[L]$ 的第 n_c 个

特征值作为比值 $c_k k/c_k^0$ 的函数来检查. 在正确的比值处第 n_c 个特征值将达到最小. 这一方法被成功地应用到芘和蒽的混合物.

M. Sjostrom 等应用偏最小二乘法与分子萤光发射光谱相结合在波长 320 — 540nm 期间测定磺酸木质素、腐殖酸和某一含有增白剂的洗涤剂，并将测定结果与主成分分析的结果相比较，采用 16 个校正混合物通过用 7 个或 8 个因子来校正所使用的方法. 为了补偿对于混合物的响应中所存在的轻微非线性，考虑数目较大的因子是必要的. 他们用 9 个混合物来评估该方法，结果发现，对于磺酸木质素和增白剂来说，偏最小二乘法所得结果较主成分分析法的稍好一些，但对于腐殖酸来说，情况恰好相反. 总的说来，这两种方法所得结果的差别并不明显.

J. Saltiel 等采用主成分分析技术研究了在不同的激发波长和不同的 O_2 浓度条件下得到的反式 2- 苯基 -2(2- 萘基)- 乙烯的萤光光谱，从而确定在所研究的化学平衡问题中存在的物种数目.

8.2.3 诱导耦合等离子发射光谱

D.F. Wirsz 等人用目标变换因子分析技术来处理混合物的诱导耦合等离子发射光谱以完成多元素的定量分析，所用的光谱是由一个低分辨率多色器和一个 1024 元光二极管阵列检测器测量而得的，由纯标准的谱构成的检验向量被用来检验混合物中 Cr，Mg，Ni，Sr 和 Zr 的存在. 当混合物中全部元素的存在都经过确认后，根据那些成功的检验向量的目标变换因子分析所得的回归系数被用来测定在某未知混合物中的元素的浓度. 他们近期又报导了一种方法，该法用因子分析去鉴别一个诱导耦合等离子发射谱中某一未知元素显示干扰的波长，然后从数据矩阵中剔除起干扰作用的波长，再应用目标变换因子分析技术去测定要求的组分的浓度.

G.E. Bentley 等应用主成分分析技术研究多组分等离子发射谱中的非随机波动现象，确定引起变动的线性无关原因的数目以及被分析组分与这些变动因子的相关情况. 其研究目的在于鉴定响应信

号漂移的原因并挑选用于内标的元素，这些元素能最佳地对仪器的漂移予以补偿.

A. Lorber 等应用一种称为内参投影法的技术通过诱导耦合等离子发射光谱数据来测定硝酸双氧铀溶液中 20 种金属的浓度. 为了对所测定的这些金属离子进行漂移校正，在 8 个波长通道对背景信号进行监测，他们检验和发现了合适的工作范围. 内参投影法可在这些范围中对等离子体的入射波功率变化、对喷雾压力的变化、对冷却气体流速的变化和铀浓度的变化进行补偿. 对正常仪器在 2 小时内的漂移得到非常满意的补偿，事实上，那些影响所选择的监测通道的因素同样也会在分析通道中产生影响，为了找到这些因素，他们采用因子分析技术来检测校正数据.

M.R. Ramsey 等用主成分分析法去鉴别存在于诱导耦合等离子体 / 原子发射光谱法 (ICP/AES) 中的噪声和漂移的原因. 研究表明存在着一种潜在的易犯的错误. 相关的多元素效应会导致方差的不准确分布. 只有通过对特征向量的检验和实验的模拟才能检测出这种现象. 因此，他们力劝人们在解析主成分时须倍加小心.

8.3 色　　谱

D. MacNaughtan 等人首次报导成功应用因子分析来解析色谱中的两个或更多的重迭峰，该法要求具有相同组分但不同浓度的几个混合物的色谱. 研究要求量测的精度，尤其是在时间轴上的精度要高，通过在相同的时间间隔对每一色谱进行数字化来构造数据矩阵，数据矩阵的每一行对应于一个给定的混合物，每一列则对应于一给定的洗提时间.

在他们的研究中，有一个是关于苯和高氘苯的混合物的，共记录了 4 个混合物的色谱，面积被归一化至 1 以便补偿由于在样品大小中的变动而引起的任何误差，对所得的数据进行因子分析并用重叠合法程序来加以处理，重叠合法所得的定量结果在 2% 范围内与

从完全的色谱分离所得的结果相一致.

重叠合法程序有一局限：色谱必须包含有每一个纯组分所产生的区域. 对于一个二组分体系，这一限制并不太严重，因为色谱的两个端尾符合这一准则.

J.E. Davis 及其合作者运用质谱去测定在一个单独的色谱峰的情况下的组分的数目. 在色谱峰的洗提过程中，他们记录在固定时间间隔的质谱，这样，每一次扫描记录了一个相同组分的不同组成的完整的质谱，数据矩阵由质谱强度组成，其中，每一列标明一个时间间隔，每一行标明一个给定的质核比. 然后对这一数据矩阵进行因子分析.

对二氧化碳 $^{13}C^{16}O_2$ 和 $^{12}C^{16}O_2$ 的同位素混合物以及正己烷和正戊烷的混合物，这一方法获得了成功的应用. 研究还包括了对色谱分辨率、峰高、峰宽和峰尾的差异所产生影响的探讨. 来自化学或电子原因的峰畸变、通道至通道的转移和在基线中的变化被发现对因子分析检出第二组分的能力没有产生大的影响. 噪声形成最严重的问题，有时候产生了一个不真实的组分，然而，这种情况将通过对实验谱图的直观审察而迅速加以确认.

鉴于对一个 200×5 矩阵做完全的主成分因子分析 (PFA) 只需 3 — 5 分钟，故这一途径为在一个色谱峰中检测出多于一个物种的存在提供了一种迅速和有用的方法，这种色谱峰看起来好象是由一单独物种所产生的. 与往常的重叠合法技术相反，因子分析法不要求对任何组分的色谱峰形作事先的假设，因此，这种方法可用以确认峰的纯度或用来发出色谱分离是无效的这样一类的警告. 一个与 GC-MS 体系有接口的小型计算机特别适合于这种研究. 因子分析法迅速、灵敏和可信.

S.D. Frans 等提出一种反复迭代最小二乘法用於从光电二极管阵列检测器测得的重叠液相色谱峰中解析出各单个组分的谱，重叠谱中的组分多达 7 种.

M.O. Eide 等人通过对高分辨气相色谱曲线的主成分模拟来进

明了氮作为杂质存在于该混合物中，戊烷 / 辛烷混合物的因子分析指出有 3 个而不是两个组分存在. 对其有关谱的仔细检查发现，离子源被以前测定过的硝基苯沾污.

W. Windig 等采用一种自模方法来解决一连续系列的热解质谱的曲线分辨问题. 该技术采用 FA 与方差图技术相结合，将来自时间分辨过的热解质谱数据的总离子流 (TIC) 曲线解析成化学组分曲线及其质谱，用于热解质谱数据的样品为生物聚合物 (脱氧核糖核酸、牛血清白蛋白和糖原)、道格拉斯杉木和橡胶共聚物，对这些样品解析结果所得的化学组分曲线与参照谱明显地相似. W.Windig 还和另外一些同事用自模曲线解析法去处理二环己基胺、二苯胺和二苄胺混合物和一组 A 链球菌细胞组织的时间 - 分辨质谱数据.

X.D. Liu 等将主成分分析和聚类分析与火花源质谱法相结合用于研究在金属锌、铜试样中痕量元素的分布情况，所得结果与直接成像二次离子显微技术的相一致.

R. Tsao 等以已有的热解质谱模式识别技术为基础发展了一种用以分析燃烧时汽油残存物的方法，来自烟雾中的挥发物被捕集在已粘结在铁磁丝上的活性碳上，解吸附后用热解质谱进行组分分析，然后采用因子分析及其随后的图形旋转去提取汽油蒸汽的因子谱. 该方法已用于鉴别喷涂单一和混合聚合物样品时所形成的汽雾中的汽油.

S. Koemig 等用主成分分析法从离子的相对贡献不同的系列混合质谱中重构各种同（原子）量异序离子的纯的级联质谱.

8.4.2 目标因子分析研究

E.R. Malinowski 等人指出质谱数据的目标因子分析法如何能被用来对怀疑其在有关混合物的一个系列中存在的物质作定性鉴定，他们还指出，如何用目标因子分析来获得混合物的化学组成. 这种对化合物鉴别的独特途径和接着进行的定量分析阐明了目标因子分析法在分析化学中的能力.

正如上一节所述的那样，含有相同的组分但各组分的量不同的系列混合物的质谱对于因子分析是很合适的，因为一个混合物的质谱中的每一个质谱峰的强度（高度）是每一组分贡献的线性加和

$$H(i,\alpha)=\sum_{j=1}^{n}h^0(i,j)P(j,\alpha), \tag{8.4}$$

式中，$H(i,\alpha)$ 是混合物 α 的第 i 个 m/e 峰的高度，$h^0(i,j)$ 是组分 j 的第 i 个峰每单位压力时的高度，$P(j,\alpha)$ 是混合物 α 中组分 j 在离子化室中的分压．鉴于质量鉴别能力，在离子化室中的分压与在储样器分压的比值对每一个组分来说都是不相同的．由于这些压力极低，运用 Dalton 定律，得到

$$H(i,\alpha)=\sum_{j=1}^{n}H^0(i,j)F(j,\alpha), \tag{8.5}$$

式中，

$$F(j,\alpha)=X(j,\alpha)(D(j)/P^0(j))P(\alpha); \tag{8.6}$$

这里，$H^0(i,j)$ 是在离子化室中压力为 $P^0(j)$ 时纯组分 j 的质谱中第 i 个 m/e 峰的高度；$P(\alpha)$ 是离子化室中的总压力；$D(j)$ 是质量分辨因子；$X(j,\alpha)$ 是原来样品混合物 α 中组分 j 的克分子分数．

方程式 (8.5) 显示出，纯组分的光谱高度是真实因子，在目标因子分析中可用作检验向量．式 (8.6) 则指明对应的余因子 $F(j,\alpha)$ 如何同克分子分数发生联系，通过应用一个已知组成的溶液，能得到混合物的组成而与压力量测无关，这是合理的，因为对于一个含有组分 1，2，…，n 的指定溶液来说，$F(j,\alpha)$ 余因子的比例与总压力 $P(\alpha)$ 无关．

应用质谱因子分析法对混合物作定性与定量分析的操作步骤是这样的：首先，通过分解协方差阵来推导出组分的数目，然后，采用怀疑其存在的纯组分的质谱作为目标因子分析中的检验向量来鉴别组分；最后通过加入一个已知组成的溶液的质谱到数据矩阵中并用纯组分的质谱作为真实向量进行组合目标因子分析便可获得各混合物的组成．

为了阐明目标因子分析技巧中的步骤，Malinowski 等用上面所讲的操作步骤去处理 Ritter 等人所获的质谱数据，数据矩阵由 7 个环己烷 / 己烷混合物的 18 个 m/e 值的强度构成．这一数据矩阵的抽象因子分析研究结果列于表 8.1 中．

表 8.1　对用以测定在一系列有关混合物中组分数目的质谱强度的因子分析结果

	环己烷 / 己烷			环己烷 / 己烷（不含 $m/e28$）			
n	RE	IE	IND×10^3	n	RE	IE	IND×10^3
1	1.810	0.684	50.27	1	1.812	0.685	50.35
2	0.465	0.249	18.62	2	0.134	0.071	5.36
3	0.128	0.084	8.03	3	0.106	0.070	6.65
4	0.111	0.084	12.30	4	0.092	0.070	10.25
5	0.098	0.073	24.56	5	0.072	0.061	18.08
6	0.074	0.068	73.51	6	0.058	0.054	58.18

从上表可以看出，采用 4 个或更多的特征向量时 IE 函数显示出很小或没有改善的事实以及当 n =3 时 IND 函数达到最小值的事实都是有 3 种组分存在的证据．第三种组分被怀疑为是作为沾污剂的氮气，当作为氮的特征的质谱 $m/e28$ 峰的强度被从数据矩阵中删除时，得到了表 8.1 中右半部的结果．这时，IE 和 IND 函数两者都显示出仅有两个组分决定着剩余下的质谱数据，由此便可确认氮的存在．这同 Ritter 等根据残余误差判据所得结论相一致．从表中可以见到，对完全的数据阵来说对应于 n =3 时的真实误差 (RE) 和对简化（即没有 m/e 28）后的数据阵来说对应于 n =2 时的真实误差都是 0.13 个强度单位，这较原来的研究者们所报告的误差 (±0.05) 要大的多．Malinowski 等人认为 0.13 这个值更加可信，因为是各种误差来源的混合物，而 0.05 只简单的是从实验图谱中读出质谱强度时的误差．应用简化后的矩阵和两个因子，进行目标因子检验，用纯的环己烷和纯的己烷的质谱强度作为检验向量，正如表 8.2 中所列的那样，在两种情况下，对于环己烷，预测的强度与检验向量在所期望的误差限度 0.13 内相一致．那些在检验向量中被自由浮动的物质的强度被正确地加以预测，这为它们的存在提供了更进一步

的证据.

表 8.2 得自目标检验和光谱分离的环己烷的质谱强度

m/e	Test[a]	TFA（预测）	光谱分离[b]
27	(1.8)	1.9	1.8
29	1.3	1.3	1.1
39	2.5	2.6	2.3
40	0.7	0.6	0.7
41	(7.1)	7.3	6.9
42	(3.5)	3.5	3.2
43	2.2	2.1	1.8
44	0.2	0.2	0.1
54	(0.8)	0.7	0.8
55	4.6	4.9	4.7
56	13.5	13.6	13.5
57	(1.2)	1.2	0.7
69	3.8	4.1	4.0
83	0.8	0.7	0.7
84	10.7	10.4	10.7
85	0.9	0.9	0.9
86	(0.1)	0.1	0.0

a. 括弧中的值被自由浮动（即在检验向量中留有空白）； b. 经过调节，使基峰为 13.5 而不是象原来文章中所报告的 100

表 8.3 得自目标因子分析和光谱分离的环己烷 / 己烷混合物组成

混合物 $F(j,\alpha)$			克分子分数环己烷		
(α)	环己烷	己烷	实验值	TFA（预测值）	光谱分离（预测值）
1	1.00	0.00	1.00	1.00	0.96
2	0.79	0.10	0.92	0.88	0.84
3	0.84	0.19	0.83	0.81	0.78
4	0.66	0.53	0.55	0.55[a]	0.54
5	0.34	0.76	0.23	0.30	0.30
6	0.17	0.87	0.12	0.16	0.17
7	0.00	1.00	0.00	0.00	0.01

a. 代表标准溶液

当在组合中用两个检验向量时，目标因子分析产生了完整套的 $F(j,\alpha)$ 余因子．含有 55% 摩尔环己烷的混合物 4 被认为是已知组成的标准溶液，采用式 (8.6)，根据 $F(j,\alpha)$ 余因子的值和混合物 4 的已知组成，Malinowski 等测出了列于表 8.3 中的混合物的组成．计算的组成与报导的组成之间的一致性很好的说明了原来的溶液只是粗略地配备而成的．

L.V. Vallis 等用热解质谱去分析生物化学混合物（包含有糖元、葡聚糖和牛血清白朊），他们将目标因子分析当作一种投影方法而不是旋转方法并对设计和改进这类混合物的分析提出了一些建议．

8.5 动 力 学

测定作为时间的函数的反应物种数目及它们的浓度是化学动力学的基础．在这样的研究中，因子分析已被表明是非常有用的手段，它能给出用任何其他方法所不能获取的信息，随着完善的计算机接口数据采集系统的出现，因子分析法的应用无疑将会增加，这些系统是为快速扫描波长动力学实验而开发的．

S. Ainsworth 的研究铺开了将因子分析法应用至动力学研究的道路，在他的开拓性研究工作中，他采用了 R.M. Wallace 的秩分析法来测定反应混合物中的吸光物种数目，他研究了细胞色素氧化酶的反应情况．在一定量的细胞色素 C 存在时，用二巯基丙醇先还原细胞色素氧化酶，被还原后的细胞色素氧化酶（混合后浓度为 2×10^{-6} M，pH9，室温）同氧（混合后为 3×10^{-6}M）反应，在 11 个波长、4 个特定的时间间隔用节流技术来测量吸光度，数据矩阵的秩被发现是 3．这暗示着在被还原了的细胞色素氧化酶的氧化过程中至少有 3 个物种对光谱的变化有贡献．这简单地被解释为体系包含有 A → B → C 类型的连续反应．

S. Ainsworth 还研究了氧血红朊和还原了的血红朊混合物，通过对氧血红朊稀溶液（pH9，硼酸盐缓冲）的逐步除氧来制备这种

混合物，将氧血红朊稀溶液装于密封的玻璃池中，在池的旁管中置还原剂（蒽醌 - β - 磺酸和连二亚硫酸钠）溶液，这种情况下，已还原的血红朊与氧的反应速度慢，因此，在要求用来获得一套反应混合物的吸光度读数（在连续的波长处）的期间，组分没有出现明显的变化．对数据的秩分析证明了血红朊的 4 种血红素对光吸收均有贡献．

在另一个研究中，S. Ainsworth 还认为，在一定的条件下，虽然对组分本身或任何组分的光谱没有什么事先的了解，但是，获取混合物中的一个组分的吸收光谱是可能的．当存在该组分不对总的吸收做贡献的情况时，上述可能性可以实现，这种情况往往存在于一个化学反应的最开始或最末了时刻，那时，产物或是反应物的浓度小得可以忽略，在这种情况下，吸光度矩阵的秩将减小一个单位，通过计算所有子矩阵的行列式，确定是否存在这样的情况是可能的，如果存在，那么在该组分不对吸收有贡献时，则更进一步的研究将会揭示出所研究的该组分的光谱， Ainsworth 成功地将这一技术应用于模拟的计算过的数据，这些数据涉及到吖啶橙，二碘代萤光素和罗丹明 B 在乙醇中的混合物．

如同 R.M. Wallace 一样， Ainsworth 的上述研究均以通过对子矩阵的奇异性检验来确定秩为基础，他们都采用标准偏差判据．

D. Katakis 通过检验由高斯消去法所产生的残余矩阵来推导出秩，他采用这种步骤研究了在 1M 高氯酸溶液中 Cr^{2+} 和马来酸之间的反应，所用吸光度矩阵由波长作为行、反应时间作为列来构成，计算结果发现，第一个残余矩阵小于误差矩阵，这个事实确定地指出体系中只有一个吸光组分存在，在推演反应的真实机理时，这样的信息是有价值的．

Z.Z. Hugus Jr 和 A.A. EI-Awady 研究了某些双核钴（Ⅲ）络合物的水解解聚作用，为了检验它们的详细的动力学模型，他们需要知道存在的吸光物种数，所用数据矩阵由 38 个溶液在 9 个波长处的吸光度所构成，为此，他们发展和应用了许多因子分析技术，他们

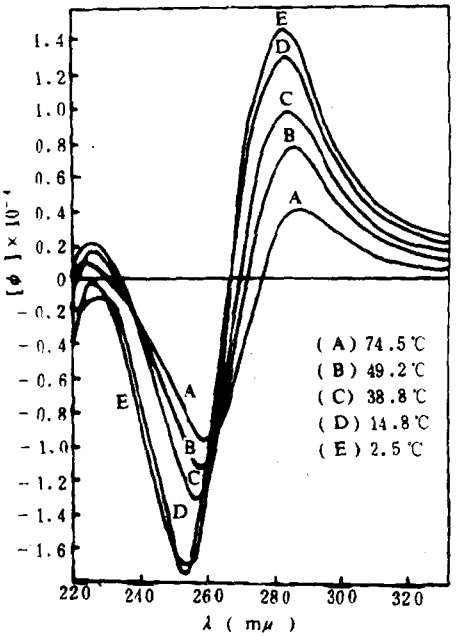


图 8.1 TMV RNA 在一系列温度下的旋光色散
(pH 7.5，在 0.04M Na^+ 中)

由于通过温度来改变组分，这种方法是相当独特的，它是以不同几何构型之间的平衡的温度依赖性为前提的. 方法富有成果，在每一种情况下，通过秩分析揭示了有两种组分存在的事实. 曾做过努力将所有的实验谱同选自实验数据本身的仅有的两个典型谱进行拟合. 1M Na^+ 时的低温谱和 0.004M Na^+ 时的高温谱被选用为两个典型向量. 之所以这样做，是因为它们代表两个极端条件，尤其是，因为增加离子强度对谱所产生的影响与降低温度所产生的影响是一样的. 采用这两个典型向量，将能在 4% 的误差范围内复原全部的数据.

由于取得了上述的成功，他们试图将这两个因子变换成 TMV RNA 分子的两个不同形式. 他们假定这两种形式是共存在化学平衡

中的单股和双股螺旋构型

$$\text{双股} \underset{}{\overset{T,I}{\rightleftharpoons}} \text{单股}$$

高温时，平衡发生偏移，致使非常少的双股构型保留下来．他们推论，单股构型的旋光性取决于温度，而双股构型的旋光性则与温度无关，两种构型的旋光性对离子强度都不敏感，然而，离子强度的变化会移动平衡．这种模型导致在实验条件范围内双股构型的百分组成和平衡常数的直接计算，在 25 ℃时， TMV RNA 看来好象有约 50% 的双股构型，而在 74.5 ℃时，双股构型则只有 3% 左右．

8.7 X 射线方法

T.H. Starks 等发展一种不加标准的定量 X 射线衍射分析法用以同时测定多相材料 (如煤灰、水泥和岩石等) 中的矿物组成和材料中各矿物的化学组成，该方法对样品中总的化学氧化物浓度数据矩阵进行主成分分析．在目标变换手续中，采用在某一给定角度下的 X 射线衍射强度量测来作为化学氧化物浓度数据矩阵的检验向量，借以在 X 射线衍射数据矩阵中发现那些对一种矿物给出独特响应的数据列．该法用以分析 3 个数据实例．第一个是煤灰试样的合成数据，涉及石英、黄铁矿、伊利水云母、高岭土、石膏矿和方解石等 6 种矿物，包括有 Si, Al, Fe, Mg, Ca, K 和 S 等 7 种元素．第二个是由石英、黄铁矿、方解石和高岭土 4 种矿物混合而得的 13 个子样品的实际数据，涉及 Al, Si, Fe, Ca 和 S 等 5 种化学元素．第三个是从磷灰石所得到的一个真实数据，涉及 Francolite，方解石，白云石和石英等 4 种矿物以及 P, Ca, F, Mg 和 Si 等 5 种化学元素．

M.F. Koenig 等用曲线拟合和因子分析去测量几种经阳极化处理过的钨薄膜的氧化物 / 金属的 X 射线光发射光谱的峰的强度比，然后用这一测量去计算在不同阳极化电压 (0.5 — 1.0V) 下所获得的氧化物薄膜的厚度．

表 8.4 目标因子分析测得的聚合物薄膜的与已知组成的比较

组成聚合物薄膜	因子分析		已知	
	PE	PMMA	PE	PMMA
聚异丁烯甲酯 (PMMA)	0.00	1.00	0.00	1.00
聚异丁烯乙酯	0.19	0.82	0.17	0.83
聚异丁烯异丁酯	0.12	0.88	0.37	0.63
聚异丁烯正丁酯	0.20	0.80	0.37	0.63
聚异丁烯十八烷基酯	0.65	0.36	0.77	0.23
聚乙烯	1.00	0.00	1.00	0.00

Cu MVV 和 Cu LMM 俄歇谱，发现在 SnO_2 掺杂的氧化镉异质结光电池的界面上存在一层金属铜. 用同样的方法，在某些溅射条件下他还在溅射涂覆铂在氧化锆陶瓷的界面上发现铂 - 锆金属互化物. 在另一个工作中，他研究了嵌入低碳钢中的铬和磷的浓度分布，同时也报导了采用二级离子质谱深度分布实验来测定在 Pb Te/Ho 掺杂的 Pb Te 薄膜中 Ho 的浓度剖析. J.H. Wandass 等通过对俄歇电子光谱深度分布的因子分析来测定在一个铜离子嵌入 50–50 Co–Ni 含量中的铜、镍和钴的浓度. J.S. Solomon 用 FA 分析二次离子的能量分布曲线，以测定在阳极氧化过的金属钽中钽在氧化钽上的定量分布. 他和他的合作者还用 FA 处理溅射深度分布时所得到的俄歇谱借以研究在 Ni/GaAs(100) 结构中氧对掺合和扩散的影响.

S. Hoffman 等用因子分析和俄歇电子能谱相结合去研究镍和镍铬铁合金的室温氧化反应，并同最小均方拟合技术做了比较，结果发现在检出化学吸附键合态方面，因子分析效果更好.

M.E. Kargencin 等应用因子分析和交互校验法去分析金属铬及其氧化物的二级离子质谱以确定在金属表面生成的氧化物的数目、种类及其浓度分布.

总的看来，在表面光谱学中应用因子分析技术主要能获得这样几种收益：①可从重叠谱中消除干扰；②可确定存在的组分数目；③可通过目标检验手续来对未知组分的存在单独地进行检验.

9 因子分析在化学基础研究中的应用

9.1 核磁共振理论研究

在这一节中，将介绍在研究取代基和溶剂对核磁共振化学偏移的影响中，应用因子分析来解决问题的一些情况.

9.1.1 质子溶剂位移的研究

NMR 谱特性 (化学偏移和偶合常数) 受到溶剂的强烈影响，这使谱的解释变得困难，不过，它却又给化学家提供了探视液体溶液状态和研究分子间相互作用的线索. J. Homer 对被认为影响溶剂位移的各种因素已进行过评述，他并已推导过许多用于这些因素的理论表达式，且已努力想通过对溶剂、溶质和其他实验变量的审慎选择来使这些因素被彼此分离. 遗憾的是，所有他的这些努力只取得了有限的成功，主要是因为不可能找到这样的实验条件，使得全部因素 (而不是一种) 的影响都维持不变. 在 NMR 中溶剂位移研究的最终目标是用最小变量集来解释其行为，鉴于当时的有关理论只能成功地进行定性的或是半定量的解释，作为理论基础的模型也因为包含有许多粗略的近似而显得可疑，于是 P.H. Weiner 和 E.R. Malinowski 等人采用因子分析技术对这方面的问题进行了一系列的研究，解决了某些简单的极性溶质 (如甲烷及其取代物) 的质子溶剂位移. 为了系统地、逐步地描述他们所采用的复杂的步骤，下面将按照他们的研究顺序来展开讨论.

1. 关键溶剂

1970 年， P.H. Weiner 等发表了第一例用目标因子分析进行质子 NMR 溶剂位移研究的结果，目的是要发展一种数学技术，以便

为 3 个列的随意组合可能不决定因子空间. 一个给定的组合可能包含不涉及一个特殊的溶质 - 溶剂相互作用的数据. 例如, 如果氢键合是一种重要的相互作用, 则至少 3 个轴中的一个说明这个因子, 这一点是绝对必要的.

乙腈具有大的偶极矩和 π 电子, 二溴甲烷具有大的极性和相当大的四极矩, 四氯化碳是非极性的且含有大体积的氯原子, 由于相信这些特性将适合於说明在因子空间中涉及到的全部可能的溶质 - 溶剂相互作用, Weiner 等把它们选择为典型溶剂的一个关键集, 事实上, 在所有的典型列向量组合被进行目标检验后, 这 3 个溶剂因子被发现提供最佳的数据复原. 对于这种选择, 虽然不存在什么唯一性, 但是, 其他的组合不旋转该因子空间, 如二氯甲烷、氯仿和四氯化碳就不能满意地复原数据. 很明显, 该空间中的重要溶质 - 溶剂因子中至少有一个不能被这组溶剂充分地加以描述.

对 3 个数据列的一个同时组合变换得到下列类型的方程式

$$\left.\begin{aligned}
\delta(\mu,\mathrm{CH_3CN}) &= 1.002f_1 - 0.004f_2 + 0.002f_3,\\
\delta(\mu,\mathrm{CH_2Cl_2}) &= 0.081f_1 + 0.7150f_2 + 0.207f_3,\\
\delta(\mu,\mathrm{CHCl_3}) &= -0.046f_1 + 0.817f_2 + 0.230f_3,\\
\delta(\mu,\mathrm{CCl_4}) &= -0.002f_1 + 1.004f_2 - 0.002f_3,\\
\delta(\mu,\mathrm{CS_2}) &= 0.006f_1 + 1.128f_2 - 0.139f_3,\\
\delta(\mu,\mathrm{CH_2Br_2}) &= -0.001f_1 + 0.009f_2 + 0.992f_3,\\
\delta(\mu,\mathrm{CHBr_3}) &= -0.256f_1 - 0.053f_2 + 1.312f_3,\\
\delta(\mu,\mathrm{CH_3I}) &= 0.561f_1 - 0.224f_2 + 0.653f_3,\\
\delta(\mu,\mathrm{CH_2I_2}) &= -0.109f_1 - 1.193f_2 + 2.295f_3.
\end{aligned}\right\} \qquad (9.1)$$

式中, $\delta(\mu,v)$ 代表溶质 μ 在溶剂 v 中的化学偏移, $f_1=\delta(\mu,\mathrm{CH_3CN})$, $f_2=\delta(\mu,\mathrm{CCl_4})$, $f_3=\delta(\mu,\mathrm{CH_2Br_2})$. 从某溶质在这 3 个关键溶剂中被测量所得的位移, 这些方程式便能预测该溶质在一给定溶剂中的化学位移, 由于实验误差和计算机的取舍, 上面为乙腈、二溴甲烷和四氯化碳各自所列出的方程式中都显示出 3 个限定的系数, 其中两个近似于 0 , 一个近似于 1. 通过检验为某些溶质预测得到的位

移值可进一步证实这些方程式的有效性和合理性，这些被检验的溶质是有意从因子分析流程中被略掉的. 例如，通过测量 CH_2Cl_2 在 3 个关键溶剂中的位移，可用式 (9.1) 来预测二氯甲烷在其他溶剂中的位移. CH_2Cl_2 在上述 3 个关键溶剂 (CH_3CN, CCl_4 和 CH_2Br_2) 中的位移分别为 326.9， 317.1 和 321.2 Hz，将这些值代入式 (9.1) 中对应于 CH_3I 溶剂的方程式进行计算得

$$\begin{aligned}&\delta(CH_2Cl_2, CH_3I)\\&\quad = (0.561) \times (326.9) - (0.224) \times (317.1) + (0.653) \times (321.2)\\&\quad = 322.2\text{Hz},\end{aligned}$$

这个值表示 CH_2Cl_2 在溶剂 CH_3I 中的质子位移，与测得的值 322.5 Hz 在实验误差范围 (± 0.5 Hz) 内吻合. 式 (9.1) 的精度通过检验那些前面提到过的有目的地搁置于因子分析流程之外的 5 种溶质在多种溶剂中的化学位移便能得到更充分的证明. 这种检验的结果列于表 9.2 中. 此外，式 (9.1) 还被用来预测取代某些非甲烷类化合物，

表 9.2　5 种取代甲烷基化学位移的实验值和计算值的比较

溶剂	溶质									
	CH_4		CH_3CN		CH_2ClCN		CH_2Cl_2		$CHClBr_2$	
	exp	pred	exp	pred	exp	pred	exp	pred	exp	pred
CH_3CN	12.1		117.6		256.8		326.9		449.7	
CH_2Cl_2	12.1	13.7	118.0	119.4	248.2	247.5	319.8	319.5	434.5	434.0
$CHCl_3$	12.7	13.8	120.0	118.8	246.1	245.7	317.4	317.9	432.1	431.0
CCl_4	13.8		117.4		244.2		317.1		430.2	
CS_2	13.3	13.6	114.8	116.1	242.8	241.9	313.9	315.0	427.9	427.5
CH_2Br_2	13.8		122.7		252.3		321.2		435.4	
$CHBr_3$	15.2	14.3	127.3	124.6	253.0	252.3	321.8	320.9	433.0	433.3
CH_3I	12.9	12.7	122.9	119.8	255.3	254.2	322.5	322.2	439.6	440.4
CH_2I_2	15.1	13.8	128.8	128.1	257.3	258.5	323.5	321.7	434.9	434.9
数据的实验变程	3.0		14.0		15.0		13.0		21.8	
平均误差	0.9		1.7		0.7		0.8		0.5	

表中值均以 Hz 为单位， 60 MHz 处，以 TMS 为内标

通过测量各种取代非甲烷类化合物在 3 个关键溶剂中的位移，便可用式 (9.1) 来完成这种预测，某些预测结果列于表 9.3 中，从表 9.3 中可以见到，即使因子的真实性质未知，通过因子分析去预测溶剂位移仍然是可能的.

表 9.3 某些非甲烷类溶质化学位移的实验值与预测值的比较

溶剂	溶质											
	CH_2ClCCl_3		$CHCl_2CCl_3$		CH_3CHBr_2		$(CH_3)_2CHBr$		Acetone		Benzene	
	exp	pred	exp	pred	exp	pred	exp	pred	exp	pred	exp	pred
CH_3CN	269.0		395.4		361.7		262.3		124.5		442.7	
CCl_4	255.8		363.0		347.3		252.0		125.4		435.9	
CH_2Br_2	261.8		372.7		354.0		258.7		128.8		441.2	
$CHCl_3$	257.2	256.8	366.3	364.1	350.6	348.6	257.3	253.3	129.7	126.4	441.1	437.3
CS_2	254.1	253.8	362.1	360.0	345.1	344.8	248.8	249.8	122.4	124.3	433.4	433.0
CH_3I	263.0	264.6	377.2	384.2	354.4	356.4	256.7	259.7	126.1	125.9	439.1	438.9
数据的实验变程	14.9		33.3		16.5		13.5		7.3		9.3	
预测值平均误差	0.8		4.0		1.6		2.7		1.3		1.4	

2. *理论考虑*

在最初因子分析成功的基础上， P.H. Weiner 等从理论方面来继续他们的研究工作. A.D. Buckingham 等人于 1960 年不经证明地提出：溶剂位移可表达成项的一个线性加和

$$\begin{aligned}\delta(\mu,v) = \ & \delta(\mu,\text{gas}) + \sigma_b(v) + \sigma_a(v) + \sigma_w(\mu,v) \\ & + \sigma_E(\mu,v) + \sigma_H(\mu,v) + \cdots,\end{aligned} \tag{9.2}$$

式中， $\delta(\mu,v)$ 是溶质 μ 在溶剂 v 中的化学位移； $\delta(\mu,\text{gas})$ 是溶质的气相位移； $\sigma_b(v)$ 是溶剂整体磁化率所引起的位移； $\sigma_a(v)$ 是由溶剂的各向异性引起的溶剂位移； $\sigma_w(\mu,v)$ 是溶剂和溶质之间的 van der Waals 扩散相互作用； $\sigma_E(\mu,v)$ 是溶剂和溶质之间的反应场相互作用； $\sigma_H(\mu,v)$ 是氢键引起的位移.

因子分析已为溶剂位移必定是一个线性加和且每一项必定是溶质和溶剂参数的积函数这一结论提供了证据，由于没有被研究的溶

质和溶剂形成强的氢键，所以，氢键作用应予忽略．这样，式 (9.2) 可写成

$$\begin{aligned}\delta(\mu,v) = \quad & \delta(\mu,\text{gas})\cdot 1 + 1\cdot\sigma_b(v) + 1\cdot\sigma_a(v) \\ & +\sigma_w(\mu)\cdot\sigma_w(v) + \sigma_E(\mu)\cdot\sigma_E(v), \end{aligned} \tag{9.3}$$

式 (9.3) 包含有 5 个因子，而因子分析却清楚地指明只有 3 个因子起作用．要解答这一困惑的问题就要立足于这样的事实：化学位移是以溶解在相同的溶剂中的痕量 TMS 为参照的，作为内标的 TMS 也与溶剂接触并产生一个溶剂位移

$$\begin{aligned}\delta(\text{TMS},v) = \quad & \delta(\text{TMS},\text{gas})\cdot 1 + 1\cdot\sigma_b(v) + 1\cdot\sigma_a(v) \\ & +\sigma_w(\text{TMS})\cdot\sigma_w(v) + \sigma_E(\text{TMS})\cdot\sigma_E(v), \end{aligned} \tag{9.4}$$

实际上，实验位移 ($\delta^{\text{TMS}}(\mu,v)$) 描述了式 (9.3) 和 (9.4) 之间的差异

$$\begin{aligned}\delta^{\text{TMS}}(\mu,v) = \quad & \delta^{\text{TMS,g}}(\mu,\text{gas})\cdot 1 + [\sigma_w(\mu) - \sigma_w(\text{TMS})]\cdot\sigma_w(v) \\ & +[\sigma_E(\mu) - \sigma_E(\text{TMS})]\cdot\sigma_E(v), \end{aligned} \tag{9.5}$$

式中，$\delta^{\text{TMS,g}}(\mu,\text{gas}) = \delta(\mu,\text{gas}) - \delta(\text{TMS},\text{gas})$，是相对于 TMS 气相位移的溶质气相位移．式 (9.5) 包括了 3 个项的加和，每一个项都是溶质和溶剂贡献的积，这一表达式预示着因子空间是三维的，这与因子分析的结果完全一致．

依照式 (9.5)，溶质的气相位移应该是一个基本因子．根据前面章节中已介绍过的目标因子分析理论可知，通过目标变换途径，每一个有疑问的因子都可单独地被加以鉴别，所以，只凭籍其自身的贡献，气相位移便可被加以检验而不一定要去构思什么模型或做什么与其他两个因子有关的详细说明．相对于气相 TMS，对溶质气相位移的变换是成功的 (见表 9.4)，这清楚地证明了气相位移是一个真实的基本因子．

我们已经知道，在进行目标检验时，不必要求检验向量是完整的．目标因子分析的一个附带优点就是它能预测目标中的那些自由浮动点．在表 9.4 中，也列出了被自由浮动的 CH_3Cl, CH_2I_2, CHI_3

和 CH_2ClBr 的气相位移预测值. 尽管两个其他未知的因子同时在起作用，但由此也可见到，通过一个单个成功的目标变换如何能获得有价值的信息.

表 9.4 气相化学位移 * 作为一个溶质因子的检验 (用 3 个因子)

溶 质	检验值	预测值	偏 差
CH_3Cl		168.2	
$CHCl_3$	427.3	427.1	-0.2
CH_3Br	146.9	147.1	0.2
CH_2Br_2	285.0	285.5	0.5
$CHBr_3$	406.9	406.8	-0.1
CH_3I	119.0	118.5	0.5
CH_2I_2		227.6	
CHI_3		301.5	
CH_2ClBr		297.7	

* 单位为 Hz，在 60 MHz 处，相对于气相 TMS

在上面研究的基础上，P.H. Weiner 等继续运用因子分析技术去鉴别溶剂对非极性溶质质子位移的影响. 为了简化问题，他们研究了由非极性溶质组成的一个子空间，这时，不存在反应场 $(\sigma_E(\mu, v))$，因为非极性溶质缺乏产生一个反应场所必须的永久性电偶极矩. 溶解于 22 种溶剂中的 6 种非极性溶质的质子位移值在传感温度为 39±1°C 处被测量，用六甲基乙硅醚 (HMD) 为外部标准，所有数据均对于整体磁化率进行了校正. 数据的实验精确度 (了解这一点对合适的因子分析是至关重要的) 被估算为 ±0.5 Hz. 全部量测数据均列于表 9.5 中.

对表 9.5 中所列的数据矩阵进行因子分析，在用相关阵或协方差阵的情况下，都发现有 3 个特征向量复盖溶剂效应空间并在实验误差范围 (±0.5 Hz) 内复原全部数据、这说明了仅有 3 个基本因子，这同上面的研究结果是相一致的. 这 3 个基本因子是气相位移、van der Waals 效应和溶剂各向异性. 为了研究这 3 个基本因子，再简单地回顾一下 Buckingham 等的工作. Buckingham 等人认为，溶质化学位移是各种贡献的线性加和，据此，Malinowski 等人认为 Buck-

表 9.5 相对于外部标准 HMD 的非极性溶质的化学位移 (Hz)

溶剂溶质	CH_4	CH_3CH_3	Neo-C_5H_{12}	c-C_6H_{12}	c-C_8H_{16}	TMS
CH_2Cl_2	10.7	49.1	53.9	83.8	90.5	-1.4
$CHCl_3$	14.6	52.8	57.7	87.0	93.7	1.9
CCl_4	16.8	55.0	58.7	87.5	94.6	3.0
CH_2Br_2	20.6	57.4	61.0	90.2	96.7	6.8
$CHBr_3$	26.0	62.6	65.4	93.7	100.2	10.6
CH_3I	18.3	55.0	59.6	89.9	95.4	5.4
CH_2I_2	35.0	69.8	73.6	101.3	107.6	19.9
CH_2BrCl	16.6	54.2	58.1	87.7	94.6	3.7
$CHBrCl_2$	19.5	56.2	60.5	90.1	96.5	5.7
$CBrCl_3$	22.8	59.4	63.2	93.0	99.1	8.0
CH_3CCl_3	16.3	54.6	58.2	88.1	95.1	3.1
CH_2ClCCl_3	16.0	53.6	57.6	87.3	93.8	2.5
$CHCl_2CCl_3$	16.2	54.1	57.9	87.2	94.0	2.8
$CHCl_2CHCl_2$	15.4	53.3	56.8	86.9	93.2	2.3
CS_2	24.3	62.1	64.6	95.2	101.1	11.0
C_6H_6	-17.9	20.5	27.0	56.4	63.7	-27.0
CH_3CN	14.0	52.5	57.7	89.7	94.2	1.9
$(CH_3)_2CO$	1.1	41.5	46.1	77.0	83.4	-8.9
$(CH_3)_2SO$	18.2	55.1	59.6	88.7	95.1	5.3
C_6H_{12}	8.5	48.5	52.1	80.8	90.5	-2.6
C_8H_{16}	10.9	49.9	53.7	84.9	92.2	-0.5
C_6F_6	-20.9	19.7	24.5	55.6	60.9	-30.3

NMR 分光计在 60 MHz 进行操作.

ingham 等的方程式可适当地表示为

$$\delta^{\mathrm{HMD},X}(\mu,v)=\delta^{\mathrm{HMD},X}(\mu,\mathrm{gas})\cdot 1+\sigma_w(\mu)\cdot\sigma_w(v)+1\cdot\sigma_a(v),\quad(9.6)$$

式中, $\delta^{\mathrm{HMD},X}(\mu,v)$ 是溶质 μ 在溶剂 v 中的相对于外部标准 HMD 的化学位移, $\delta^{\mathrm{HMD},X}(\mu,\mathrm{gas})$ 是溶质 μ 相对于外部标准 HMD 的气相位移值 (均对整体磁化率作了校正), $\sigma_w(\mu)$ 和 $\sigma_w(v)$ 分别为与溶质中 μ 和溶剂 v 有关的 van der Waals 效应, $\sigma_a(v)$ 是由溶剂 v 引起的各向异性位移. 依照因子分析, 这一方程式的每一项必须是溶质和溶剂参数的积函数, 对于气相项, 这一约束能得到满足, 因为气相位移只是被研究的溶质的一个函数, 与溶剂无关, 与此相似, 溶剂各向异性项完全取决于溶剂而与溶质无关, 因此, 这两个项均

可表示成积函数，前者的溶剂部分为 1，后者的溶质部分为 1. 有些研究人员已提出，van der Waals 屏蔽可表达成一个积函数，与式 (9.6) 一致. 为了详细检验式 (9.6)，需要对式右边的所有 3 个项有一个精确的计算. 由于 van der Waals 项和溶剂各向异性项都是理论项，这类信息不易获得. 对于这些量，只能对非常有限种的溶剂做粗略的估算，如苯和二硫化碳等. J.C. Schug 对这两种溶剂的各向异性做了理论的估算，分别为 -30 Hz 和 $+18.1$ Hz. 另外，有人则提出分别为 -35 Hz 和 $+7$ Hz.

Weiner 和 Malinowski 用下面的方式解决这一错综复杂的问题. 以甲烷为溶质，式 (9.6) 可写成

$$\begin{aligned}&\delta^{\mathrm{HMD},X}(\mathrm{CH_4},v)\\&\quad=\delta^{\mathrm{HMD},X}(\mathrm{CH_4},\mathrm{gas})+\sigma_w(\mathrm{CH_4})\cdot\sigma_w(v)+1\cdot\sigma_a(v),\end{aligned}\tag{9.7}$$

解出这一方程中的 $\sigma_w(v)$ 并代入式 (9.6) 得

$$\begin{aligned}&\delta^{\mathrm{HMD},X}(\mu,v)\\&\quad=\delta^{\mathrm{HMD},X}(\mu,\mathrm{gas})\cdot 1+(\sigma_w(\mu)/\sigma_w(\mathrm{CH_4}))\cdot\delta^{\mathrm{CH_4\cdot g}}(\mathrm{CH_4},v)\\&\quad+(1-\sigma_w(\mu)/\sigma_w(\mathrm{CH_4}))\cdot\sigma_a(v).\end{aligned}\tag{9.8}$$

式中，$\delta^{\mathrm{CH_4\cdot g}}(\mathrm{CH_4},v)=\delta^{\mathrm{HMD},X}(\mathrm{CH_4},v)-\delta^{\mathrm{HMD},X}(\mathrm{CH_4},g)$，代表甲烷在溶剂 v 中相对于 $\mathrm{CH_4}$ 气相的气相至溶液的位移. 与式 (9.6) 比较，这一表达式的优点是它与 van der Waals 模型无关，不必说明 $\sigma_w(v)$. 用一个新的因子 (甲烷的气相向溶液的位移) 来取代那个溶剂因子. 这个步骤阐述了基本因子之间的复杂的相互依赖性，也阐明了这样的事实：存在许多表达因子空间的不同方式，每种方式同样都是合理的.

从溶剂检验的观点来看，单位向量和甲烷的气相至溶液位移向量都容易构造，因为全部必须的数据都可获得. 另一方面，溶剂各向异性的问题就不那么简单. 为解决这一问题，Weiner 和 Malinowski 把四氯化碳的溶剂各向异性定为 0(因为它是非极性的和对称的)，然后，系统地变化苯和二硫化碳的溶剂各向异性，使与各种理论估

算值相一致. 因为存在 3 个因子，而且也因为仅有 3 个其溶剂各向异性只能获得粗略估算的溶剂，所以，要想单从目标检验的结果得出这些估算中究竟那一个最好的结论是不可能的. 之所以出现这种情况，是因为对于一个 3 因子空间，任何仅含 3 个已定义过的数的检验因子都产生一个完美的拟合. 在这种情况下，因为任何 3 个随机数都会完美地拟合，所以，必须推导出另外的判据.

式 (9.8) 为选择适当的各向异性检验向量提供了必要的判据. 根据该方程式，如果溶剂各向异性向量正确的话，则溶剂单位检验向量的系数应与溶质的气相化学位移相对应，且式中最后两个系数的加和应等于 1. 如果对于 3 个溶剂检验向量 (1, 甲烷气相位移，溶剂各向异性) 进行同时目标变换，则可通过将预测得的载荷 (即式中的溶质系数) 同上面所讲到的判据加以比较便可推出最佳的各向异性检验向量. 包含这 3 个溶剂因子的目标检验结果列于表 9.6 中.

表 9.6　溶剂余因子目标检验

溶　剂	单位		$\delta^{CH_4,g}(CH_4,v)$		$\sigma_a(v)$		$\sigma_w(v)$	
	检验	预测	检验	预测	检验	预测	检验	预测
CH_4	1.0	1.014	25.2	25.4	0.0	0.0	0.245	0.273
$CHCl_3$	1.0	1.009	23.0	23.1	-	0.8	0.238	0.242
$CHBr_3$	1.0	1.002	34.4	34.5	-	4.9	0.323	0.310
CH_2I_2	1.0	0.974	43.4	43.2	-	14.8	0.301	0.286
CH_2ClCCl_3	1.0	1.007	24.4	24.5	-	-0.1	-	0.264
C_6H_{12}	1.0	1.001	16.9	16.9	-	-1.3	-	0.203
CS_2	1.0	0.992	32.7	32.6	9.0	9.0	-	0.246
$(CH_3)_2CO$	1.0	1.003	9.5	9.5	-	-5.9	0.175	0.180
C_6H_6	1.0	0.977	-9.5	-9.7	-24.0	-24.0	0.199	0.186
C_6F_6	1.0	1.002	-12.5	-12.4	-	-27.6	-	0.200

为了得到最佳的溶剂各向异性向量，用下面的方式来生成并检验各种各样的检验向量. 四氯化碳的溶剂各向异性被设定为 0，苯和二硫化碳的估计值的所有组合连同 1 和甲烷的气相至溶液的位移一起被应用在组合步骤中，这一手续所产生的溶质系数与指定的判据没有产生满意的一致性.

继续进行探索，再次将四氯化碳的溶剂各向异性设置为 0，同时，苯的各向异性值在 −40 至 −20 Hz 之间、二硫化碳的则在 +5 至

+20 Hz 之间系统地变更. 对每一次组合，都将预测所得的溶质系数和提出了的判据做比较. 当 σ_a(苯) = −24.0 Hz 和 $\sigma_a(CS_2)$ = +9.0 Hz 时，得到最佳的拟合，并导出下面的方程式

$$\delta^{HMD,X}(CH_4, v) = -8.3f_1 + 0.99f_2 + 0.01f_3,$$
$$\delta^{HMD,X}(CH_3CH_3, v) = 35.9f_1 + 0.72f_2 + 0.28f_3,$$
$$\delta^{HMD,X}(\text{neo} - C_5H_{12}, v) = 41.6f_1 + 0.64f_2 + 0.32f_3,$$
$$\delta^{HMD,X}(C_6H_{12}, v) = 75.8f_1 + 0.49f_2 + 0.55f_3,$$
$$\delta^{HMD,X}(C_8H_{16}, v) = 83.8f_1 + 0.38f_2 + 0.63f_3,$$
$$\delta^{HMD,X}(TMS, v) = 12.8f_1 + 0.63f_2 + 0.35f_3.$$

这里 $f_1 = 1$， f_2 是甲烷气相至溶液的位移， f_3 是溶剂的各向异性，可将 f_1 项的系数同实验的气相位移 (相对于外部标准 HMD) 值 (甲烷为 −8.4 Hz 、乙烷为 35.5 Hz 、新戊烷为 42.1 Hz 、环已烷为 75.6 Hz 、四甲硅烷为 16.4 Hz) 进行比较. 除四甲硅烷外，上面方程式中所示的预测气相位移值与实验值都有良好的一致性，而且，对于所列出的每一个方程式来说，右边最后两项的溶质系数的加和都接近 1. 由此可见，因为关系到溶质系数的两个判据已经得到满足，溶剂各向异性检验向量应该被认为是可信的，“最佳” 的各向异性检验向量已列于表 9.6 中，对于那些被自由浮动的溶剂的预测溶剂各向异性也列于该表中. 这些预测代表了溶剂各向异性的第一个经验估计，卤代溶剂，如二碘甲烷、三溴甲烷等，被预测具有大的溶剂各向异性. 对于卤代甲烷类的溶剂各向异性理论的发展，这些值可作为一个引导.

在获得一套满意的溶剂各向异性之后， Weiner 和 Malinowski 继续进行研究工作，试图阐明 van der Waals 因子. van der Waals 位移的任何模型必须满足 4 个判据. 第一个判据涉及到 3 个溶剂检验因子在实验误差范围内预测数据矩阵的总的能力，后 3 个判据则涉及到式 (9.6) 中的溶质系数. 如果 van der Waals 模型正确，则单位检验因子的溶质系数应该与溶质气相化学位移相对应，各向异性位移项的溶质系数应该为 1， van der Waals 项的溶质系数应与模型相

一致. 已提出几种理论模型来解释涉及非极性溶质的 van der Waals 效应. B. Linder 等将溶质当作一个振荡偶极，将溶剂当作一个介电连续统来处理，与 van der Waals 效应有关的屏蔽常数 σ_w 与由溶质振荡偶极矩所产生的振荡电场 E 的平方成正比

$$\sigma_w = \Phi E^2, \tag{9.9}$$

式中，常数 Φ 业经估算为 $-1 \times 10^{-12}\mathrm{cm}^4/\mathrm{esu}^2 \cdot \mathrm{ppm}$ ，振荡电场 E 则为

$$E^2 = (3/4)\mathrm{hg}(v_\alpha v_i)/(v_\alpha + v_i), \tag{9.10}$$

这里， h 是普朗克常数， v_α 和 v_i 分别为溶剂和溶质的振荡偶极矩的频率， g 则被定义为

$$g = ((2n^2 - 2)/(2n^2 + 1)) \times (1/a_i^3), \tag{9.11}$$

式中， n 是溶剂的折光指数， a_i 是溶质的翁隆格分子半径. 振荡频率与电子云分布半径 r_j 和分子极性 α 有关

$$v = 2/3(e^2/h\alpha) \cdot \sum\langle r_j^2\rangle. \tag{9.12}$$

H.J. Bernstein 等人采用了溶质 - 溶剂相互作用的一个维里扩展，这两种理论有许多共同的地方，不过细节处理不一样. 由于 Linder 的模型相对来说要简单一些，所以， Weiner 等人检验了 Linder 的连续统模型. 根据上面所述 Linder 的理论，在除去一个不必要的近似后，对 van der Waals 贡献，他们得到下面的表达式

$$\begin{aligned}&\sigma_w(\mu, v)\\ &\quad = [k/V_\mu]_\mu \left[\left(\left(n^2+2\right)/\left(2n^2+1\right)\right) \cdot \left(\sum\langle r_j^2\rangle/V\right)\right]_v,\end{aligned} \tag{9.13}$$

式中， k 是一个常数， V_μ 和 V 分别为溶质和溶剂的克分子体积. r_j 是溶剂分子中的一个电子的电子云分布半径，加和项包括分子的全部电子， $\sum\langle r_j^2\rangle$ 的计算采用 P.G. Maslov 的重叠关系. 根据式 (9.13) ，溶剂的 van der Waals 项的值可根据文献中的信息加以估

算. 应用目标变换，得到表 9.6 中最右列的结果. 考虑到模型的不够成熟，这种拟合却是出人意料的好.

3. 基本溶剂因子的组合

在目标因子分析中，可以单独地检验每一个基本因子，所以，人们可以将注意力集中在任何一个基本因子上面. 对理论模型最后检验要求对 3 个基本的溶剂因子进行同时的变换，与式 (9.7) 相一致，这 3 个溶剂因子是：单位因子、$\sigma_w(v)$ 和 $\sigma_a(v)$. 这 3 个检验向量的同时变换产生以下的方程式

$$\left.\begin{aligned}
&\delta^{\mathrm{HMD,X}}(\mathrm{CH_4}, v) = -10.4(1) + 103.8\sigma_w(v) + 1.10\sigma_a(v),\\
&\delta^{\mathrm{HMD,X}}(\mathrm{CH_3CH_3}, v) = 34.3(1) + 72.5\sigma_w(v) + 1.07\sigma_a(v),\\
&\delta^{\mathrm{HMD,X}}(\mathrm{neo-C_5H_{12}}, v) = 40.2(1) + 65.0\sigma_w(v) + 1.03\sigma_a(v),\\
&\delta^{\mathrm{HMD,x}}(\mathrm{C_6H_{12}}, v) = 74.7(1) + 45.4\sigma_w(v) + 1.04\sigma_a(v),\\
&\delta^{\mathrm{HMD,X}}(\mathrm{C_8H_{16}}, v) = 82.9(1) + 39.0\sigma_w(v) + 1.06\sigma_a(v),\\
&\delta^{\mathrm{HMD,X}}(\mathrm{TMS}, v) = -14.2(1) + 63.5\sigma_w(v) + 1.04\sigma_a(v).
\end{aligned}\right\} \quad (9.14)$$

这些方程式在实验误差内复原测量的位移值，对上述方程式中载荷 (即溶质系数) 的观察可证实这种分析方法的合理性，单位因子的系数与测量所得的气相位移值 (甲烷为 −8.4 Hz，乙烷为 36.6 Hz，新戊烷 42.1 Hz，环己烷 75.6 Hz，四甲硅烷 −16.4 Hz) 有相当好的一致性. 依照式 (9.6)，$\sigma_a(v)$ 的系数应该为 1，式 (9.14) 中该项的系数都很接近于 1. 最后，上述各方程式中 van der Waals 项的系数对式 (9.13) 的溶质部分作图显示出一个线性的趋势，但有一些点稍微离散，这种离散点的出现也许是由于氢原子在不同的溶质中相对于分子的中心来讲占据的位置不相同所造成的，因子分析研究所用的 van der Waals 模型不考虑这样的位置因素.

4. 概括

因子分析的最终目的是转变一个实验点矩阵成为一套能揭示所观察到的现象的本质的方程式，上述研究清楚地表明了这样的目的是可以达到的. 在这里，我们已经看到因子分析如何被用来解决引起质子 NMR 溶剂位移的重要相互作用，只有在出现 TFA 之后，才

在实验技术上的困难，他们不使用纯溶剂而使用含 20% (体积) TMS 的溶剂. 在 TMS 核素情况下，位移被用纯的 TMS 作对照. 对于环己烷核素，溶液 (除含 20 % (体积) 的 TMS 外尚含 2% (体积) 的环己烷) 和位移被同 TMS 中含 2%(体积) 的环己烷进行对照.

对 4 套溶剂进行主因子分析，第一套由环己烷、苯、邻 - 二氯苯、和 1,2,3- 三氯苯组成；第二套由 14 个溶剂组成；第三套有 15 个溶剂 (在第二套中再加入六氟苯组成)；第四套则包含有全部 38 种溶剂 (因为只对 4 种核素，才有完整的数据可利用，故排除了 ^{29}Si 位移). 最前面的两个特征向量在实验误差内复原了第一、第二和第三套数据，对于第四套，则要求三个因子，这暗示着在这个大的系统中存在着诸如溶质溶剂之间的一个偶极 — 诱导偶矩相互作用的一种另外的影响. 为了确定到底是哪一种或哪些溶剂造成这种另外的影响，将 "可疑" 的溶剂加入到第三套中去并进行因子分析. 一系列的这类检查揭示出六氟苯和二硫化碳是神秘的第三个因子的主要贡献者，溶剂的极性被认为不是第三个因子的贡献者，因为具有较高极性的丙酮和氯仿对第三个特征值没有明显的影响. 也有设想认为，也许是由于分割成溶剂 - 溶质对的状况开始被破坏，因此而产生另外一个因子.

对于第一、第二和第三套数据，一个包含有单一旋转角 θ 的二维抽象变换矩阵 $[T]$ 是合适的

$$[T] = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \tag{9.15}$$

通过以 5° 为增量旋转特征向量来检验溶剂余因子的完整范围，这一过程对单卤原子代苯和环己烷所得结果被阐明在图 9.1 中. 为利用对应于扩散相互作用和各向异性的两对溶剂 - 溶质向量来解释这个图，注意观察，正如图 9.1 中交叉线条所表示的那样，特征向量的 90° 旋转角对环己烷类产生了一个公共的因子，对苯类却产生了另一个不同的公共因子. 这一不同的公共因子被解释为是由于溶剂各向异性所造成的. 各向异性因子的溶质系数随不同的核素逐一变

B9

化：^{1}H (TMS) 为 1.04 ， ^{13}C (TMS) 为 1.43 ， ^{1}H(CHX) 为 1.26 和 ^{13}C (CHX) 为 1.11. 这些系数明显地是溶质分子的不同核的各种位置因素的定量测量. 160° 的旋转角产生了一种这样的状况，此时，卤素取代基是重要的，在这一角度，与一给定的卤代苯有关的溶剂余因子等于与相对应的卤代环己烷有关的溶剂余因子 (图 9.1). 根据 Bacon 和 Maciel 的意见，这种情况与他们的分散相互作用准则是相一致的，但与其他的分散模型不相一致.

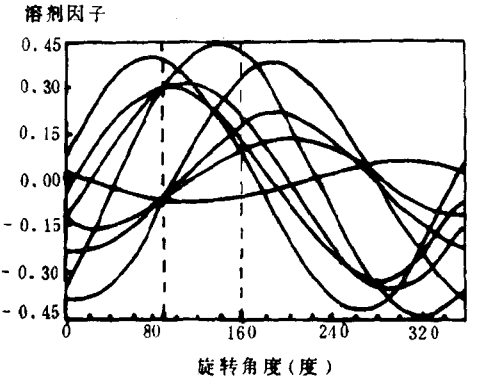


图 9.1 第三套数据溶剂余因子对正交旋转角度的依赖性 (按在 0 度处值递增顺序), 溶剂为碘代环己烷、碘代苯、溴代环己烷、溴代苯、氯代环己烷、氯代苯、环己烷和苯

9.1.3 氟溶剂位移

R.J. Abraham 等测量了 19 种刚性的非极性溶质在 8 种溶剂中的 F 位移并指出，有两种因子 (溶质气相位移和 van der Waals 效应) 对溶剂位移起作用. 当用因子分析去处理这些数据时，根据判据 IE 和 IND ， E.R. Malinowski 等找到了不是两个而是 3 个因子在起作用的证据，这第三个因子虽然小，但它被怀疑是由溶剂各向异性所造成的，尽管已知显示出大的溶剂各向异性的苯和二硫化碳并不包括在所采用的溶剂中，由于所涉及的溶剂的各向异性值未知，故无

法检验这种影响. 在另一项研究中, Malinowski 通过对气相位移进行目标检验, 确认其为一个基本因子, 检验向量和预测向量列于表 7.12 中. 同时列于该表中的理论误差也证明了气相位移是一个基本因子, 该研究尚指出, 预测所得的气相位移较测量所得的气相位移更准确 (REP= 0.05ppm, RET= 0.19ppm). 由于溶液位移较气相位移更精确, 故出现这种情况.

9.1.4 取代基对 ^{13}C 位移的影响

为探索 ^{13}C 位移的本质, K.B. Wiberg 等着眼于对卤代烃类的研究, 他们对 "差示" 位移 (即卤素取代化合物与其未被取代的母体化合物之间的位移的差值) 进行了因子分析. 所用的数据矩阵由 62 个 "差示" 位移作为列指定、四个卤素 (F, Cl, Br 和 I) 作为行指定. 分析结果得到两个因子, 目标检验产生 3 个成功的卤素性质向量, 一个单位向量 (1 1 1 1) 、一个算术级数向量 (1 2 3 4) 和一个双峰向量 (1 0 0 1). 通过在组合中应用这 3 个向量, 对涉及的每一个碳核得到 3 个有关的载荷余因子, 这 3 类不同的载荷被认为分别是由一个 "假设的" 具有相同量子数的卤素、价电子的 "自由度" 以及 "构象" 影响所造成的.

W.F. Reynolds 等对 75 种 3- 和 4- 取代 Sytrenes 的 ^{13}C 位移进行了因子分析, 涉及 15 种普通的取代基 H, $N(CH_3)_2$, NH_2, OCH_3, SCH_3, CH_3, $Si(CH_3)_3$, F, Cl, Br, $COCH_3$, CO_2CH_3, CF_3, CN 和 NO_2 . 3 个抽象因子在实验不确定性范围内说明了 "实际" 的化学位移 (相对于内标 TMS 的量测) , 不过, 仅需要 2 个因子便可说明取代基的位移 (相对于母体化合物的量测) . 在各种取代基的参数中, 目标检验指出, Taft 常数 σ_F 和 σ_R^0 是 2 个最合适的取代基因子. 对于 "实际" 位移数据来说, 单位值检验也被证明是一个有效因子, 由于 TMS 参比是随意的, 因此, 这一结论是可意料到的. 如要说明 *ipso* 碳、邻位碳和间位碳位移, 则对于每一种碳还需要另外一个因子.

H.M. Hutton 等对 4- 取代基苯酚类和 2- 硝基苯酚类的 ^{13}C 位移进行了类似上面的研究，发现对位碳的位移需 2 个因子，而 *ipso* 碳、邻位碳和间位碳则需要 3 个因子，他们同样认为 Taft 诱导和场效应参数是合适的因子.

对于 4 种不同的碳原子， Z. Urbaniak 等研究了 13 种直立键取代 4- 特丁基环己烷和 13 种平伏键取代 4- 特丁基环己烷的 ^{13}C 位移数据 (矩阵为 26×4) . 13 种取代基为 Cl, Br, I, OH, OCH_3, $OOCCH_3$, $OOCCF_3$, OTs, NH_2, $NHCH_3$, $N(CH_3)_2$, NO_2 和 CH_3 . 2 个因子占了数据总方差的 99.6% . 目标检验鉴别此 2 个目标为电负性和取代基效应.

9.1.5 其它机理研究

M. Azzaro 等对烯胺酮类的 ^{13}C 核磁共振谱进行主成分分析，借以研究取代基效应通过不饱和体系传递的情况. 他们对 3 个 SP^2 碳和两个亚甲基碳在环乙酮部分上的化学位移值进行因子分析，发现两个因子可足够占数据总方差的 93% 以上，最重要的轴 1 (79%) 对应的因子与烷基氮取代基的诱导和位阻效应有紧密的联系，但第二个因子则难以解释，可能与含氨基的功能团的 *ipso* 效应有关.

9.2 色谱性质的研究

色谱法集分离和分析于一身，是分析化学中一种非常重要且广泛应用的分析方法，不断加深对色谱过程的了解，是发展和创新色谱法的基础. 大量的研究表明，因子分析技术可被用来鉴别影响溶质 - 溶剂相互作用的基础因子，此外，也可用来对溶质和固定相的溶剂进行相似性归类. 因子分析技术的这些贡献可帮助人们更好地去认识和了解色谱过程.

9.2.1 色谱与因子分析

为什么因子分析技术也可用于研究色谱行为呢？这一小节主要

来回答和阐述这个问题. 色谱的保留值是色谱过程的很重要的数据，测量所得的保留值是所有相互作用力的一个综合体现，这些相互作用力控制着溶质通过色谱柱的运动. 色谱中普遍存在的相互作用力是由 London 色散力所引起的，这种色散力来源於诱导偶极 - 诱导偶极相互作用. 在非极性电解质的选择性滞留中，色散相互作用力起着支配作用，如果溶质或溶剂是极性的，则包含有偶极 - 诱导偶极力和偶极 - 偶极力的极性相互作用力变成主要的因素. 在某些分离中，包括诸如氢键这样的电子授体 - 受体力的相互作用力起着决定性作用，在一些特殊情况下，位阻因素表现出重要的作用.

以特定的保留体积和保留指数体系为基础的两种报告 GLC 数据的方法对因子分析都是合适的，对于色谱分布现象，校正过的特定保留体积 V_r^0 的对数与溶液的标准克分子自由能成正比，由于需要进行仔细的控制才能测量得精确的 V_r^0 值，因此，保留指数体系便成为更受欢迎的报告 GLC 数据的方法.

在保留指数法中，溶质在一个同系物系列中的保留体积被用作参比点，相对于那些标识溶质，一个溶质在这一标度上的位置被予以测量. 溶质 i 在固定相溶剂 k 上的保留指数 I_{ik} 被定义为

$$I_{ik} = 100n + 100\frac{\lg V_i - \lg V_n}{\lg V_{n+1} - \lg V_n}, \tag{9.16}$$

式中，n 和 $n+1$ 是在溶质 i 之前和之后直接被洗脱的标识溶质中的碳原子数目，V_i, V_n 和 V_{n+1} 则分别是第 i 个溶质的保留体积和两个标识溶质的保留体积. 根据 L.Rohrschneider 的热力学观点，可以期望保留指数 I_{ik} 可被表达成积项加和，因而也就可期望保留指数的矩阵具有因子分析解. 换言之，色谱的保留值问题的因子分析解具有一般的形式

$$d_{i\alpha} = \sum_{j=1}^{n} \mu_{ij} v_{j\alpha} \tag{9.17}$$

式中，$d_{i\alpha}$ 是测得的溶质 i 的色谱在溶剂 α 上的保留值，μ_{ij} 和 $v_{j\alpha}$

分别为溶质 i 和溶剂 α 的第 j 个余因子. 用矩阵形式表示, 式 (9.17) 可写成

$$[D] = [U][V], \tag{9.18}$$

式中, $[D]$ 是色谱保留值数据矩阵, $[U]$ 和 $[V]$ 分别是溶质和溶剂余因子矩阵.

9.2.2 活度系数预测和溶剂的分类

P.T. Funke 等最先将因子分析技术应用到色谱理论研究中, 他们应用主因子解去预测活度系数. 根据化学准则, 选择出 5 个典型溶质和 5 个典型溶剂的关键集, 用以预测的余因子系数被取自主因子模型而不是取自组合目标因子分析解, 对一个新溶质和 5 个新溶剂的活度系数的预测相当顺利. 虽然目标检验在这里不被采用, 但这一研究预示了目标因子分析的重要特征.

抽象因子分析已被用来鉴别在固定相中的主要的相似性和差异性, 聚类是以在主因子解中余因子的相似性为根据的. 不过, 在对 GLC 数据所进行的分类研究中, 象方差最大这样的旋转技术尚未被采用.

在抽象因子分析对溶剂分类的开创性应用研究中, S.wold 等选择了一个三因子模型作为一般性和特殊性之间的一个合理的折衷. 3 个主成分在大约 30 个 r.i. 单位内复原了数据, 第一个因子被认为是溶剂的极性引起的, 第二个因子被一个未加指明但相对恒定的溶质参数来说明其成因, 第三个因子被认为是包含在醇类溶质中的氢键相互作用所形成.

D.H. McCloskey 等、S.R. Lowry 等和 M. Chastrette 在各自的研究中采用类似的方法来对溶剂进行分类. 为了用一种易於想象的方式来描述溶剂, 采用了二因子解. 溶剂余因子 1 对溶剂余因子 2 的作图显示出溶剂的聚类, 在这样的图上, 每个点代表一个单个的溶剂. 这一方法的实用已被阐明在图 9.2 中. 图中所涉及的 225 个

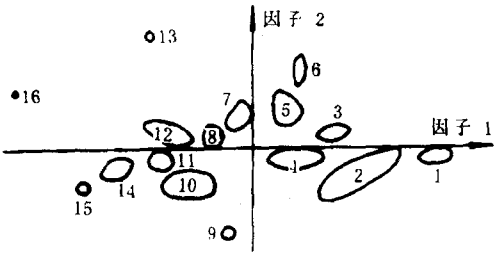


图 9.2　225 种溶剂的主余因子二维图

在每一个聚类中溶剂的类型为：1. 从配极 SP；2. 硅酮类；3. 烷基己二酸脂类；4. 酞酸脂类；5. 乌康和相类的 SP；6. 酰胺类；7. 脂肪酸乙酯类；8. Igepal 和相类的 SP；9. 氟化的 SP；10. 含氰基 SP，如，XF 1150；11. 多元醇酯类，如，乙二醇己二酸酯；12. 卡波蜡蜡和相类的 SP；13. Siponat 和 Stepan；14. 琥珀酸脂；15. 氰基乙氧丙烷和相类的 SP；16. 双甘油；(SP 代表固定相).

固定相溶剂的聚类与化学方面的理解相一致，那些其余因子远离两个轴的固定相很可能是受欢迎的. 在这种二维图的另一个应用中，M. Chastrette 提议把对溶质余因子所画的图迭加到为溶剂余因子所画的图上，这时，彼此之间相互作用最强烈的溶质 - 溶剂对在迭加后的图上往往会具有彼此靠近的点.

9.2.3　因子数目的确定及其应用

对色谱保留作用数据进行因子分析是深入了解色谱中溶质 - 溶剂相互作用的一种可行的手段，首先要碰到的是在某一色谱法体系中到底有多少种因素对色谱过程产生有意义的影响这样一个问题.

R.B. Selzer 和 D.G. Howery 对二甲醚等 18 种醚类在四氯酞酸丁酯等 25 种固定相上的保留指数进行分析，结果发现，用 5 个因子时，数据复原的平均误差是 2.8 保留指数 (RI) 单位，最大的误差是 14 个 RI 单位，当用 6 个因子时，平均误差降至 2.1 个 RI 单位，最

大的误差降至 8.2 个 RI 单位，只有 3.6% 的数据有大于 5 个 RI 单位的误差. 由于实验误差的上限为 5 个 RI 单位，因此，他们认为该醚 - 溶剂相互作用空间可以用 6 个因子来予以适当的描述. E.R. Malinowski 采用 IND 函数也得出相同的结论.

P.H. Weiner 等应用抽象因子分析中的复原步骤去研究一系列简单的溶质 (包括正己烷、环己烷、四氯化碳、氯仿等共 9 种化合物) 在 1.46(重量) 克分子的 $(CH_3CH_2)_4NBr$ 的含水电解质相上的气 - 液相色谱保留作用混合机理. 他们发现，如果只考虑饱和烷烃溶质，则仅需 1 个因子便足够旋转其因子空间并在实验误差范围内复原数据；而如果要同时考虑极性和不饱和溶质的话，则需要 2 个因子才能在实验误差范围内复原数据. 从因子分析估算得到的因子数目与应用于保留值机理的化学模型相一致.

D.G. Howery 等运用复原步骤去加深关于溶质对因子空间复杂性产生影响的问题的了解，他们研究了当数据列被加至矩阵或从矩阵中被删去时数据复原所发生的变化，他们从一个只含有链烷溶质的数据矩阵开始，测定了当代表一个或两个新溶质的数据列被增加到数据矩阵中去时因子空间的增大情况. 秩的增大与新的溶质中功能团的复杂性有联系，例如，具有一个烷基支链的某一溶质对因子个数几乎不产生影响，而一个醇溶质却使秩增大几乎近 1 个单位. 在从数据矩阵中删去一个或多个溶质所产生影响的研究中，他们发现，对于极性溶质和具有化学独特性的溶质，往往可观察到秩有最显著减小的现象.

G. Musumarra 等应用主成分分析技术来研究碱性和中性药物的薄层色谱洗提体系的选择，所用数据涉及 55 种药物在 40 种洗提混合物中的 RF 值. C.L. Deligny 等应用三模 FA 研究正向 HPLC 中的保留时间，所用数据涉及 39 种溶质、6 种吸附剂和 2 种洗提剂. E. Fernandez -sanchez 等用因子分析去确定气相色谱的混合固定相中保留体积的线性. 他们同时还用因子分析技术去减少在气相色谱中的特征溶质的数目.

9.2.4 独特性检验与单位值检验

在试图采用空间的抽象特征向量去鉴别具有化学或物理意义的参量之前，对所进行研究的全部溶质和溶剂进行独特性检验是很有用的. 这种检验能帮助色谱工作者找出那些产生特殊的相互作用的指定，也就是说，能找出那些包含有独特因子或独特相互作用项的溶质和溶剂，这种独特的因子或独特的相互作用项对其它的溶质和溶剂来说是不起作用的. 独特性检验还可帮助找出那些包含有显著误差的溶质或溶剂的有关数据点.

在色谱保留值的研究中，独特性检验的基本思想和做法是这样的，假设溶质中的第 j 个在加和中具有独特因子，则在式 (9.17) 所示方程中，这一独特因子表现为第 $(n+1)$ 项，即对于正常数据，有

$$d_{i\alpha} = \mu_{i1}v_{1\alpha} + \cdots + \mu_{in}v_{n\alpha},$$

对于具有独特因子的数据，有

$$d_{j\alpha} = \mu_{j1}v_{1\alpha} + \cdots + \mu_{jn}v_{n\alpha} + \mu_{j,n+1}v_{n+1,\alpha},$$

显然，除了具有独特因子的那个溶质之外，第 $(n+1)$ 项对所有其它的溶质都应为零. 用于进行独特性检验的因子可以这样来组成，即对于具有独特性因子的溶质，可赋予任何一个常数 (例如 1)，而对于其它的溶质均赋予零. 如果对这一检验向量的检验结果是成功的话，则意味着该检验因子中对应于非零值元素的溶质在用以描述 $d_{j\alpha}$ 的方程式中含有一个独特的部分. 如果某一溶质具有独特因子，则或者是因在数据中对于该溶质所存在的实验误差或者是因在该溶质与全部溶剂之间存在着一个真正的独特化学相互作用而造成的. 例如，如果全部其余的溶质都是非极性时，在一个数据集中，一个极性溶质就会具有一个独特因子. 这种讨论在将原始数据集进行转置和把溶质换成溶剂时也是适用的.

根据文献介绍，产生高独特值的部分溶质是：吡啶、甲醇、异丁酸丁酯、异丁酸异丙酯、乙苯、2- 辛炔、环己烷、十八烷、二甲醚、

异丁甲醚、乙醛等；产生高独特值的部分溶剂是：二甘醇 - 琥珀酸脂、双甘油、聚苯基醚 (五环) 、碘化十六烷基、正十七烷、蔗糖八醋酸酯和四元醇等. 在大多数情况下，独特分子或者是有极性功能团或者是有小的分子质量，例如，溶剂双甘油显著的独特性预示着双甘油的氢键能力. 那些从化学观点来看是不应该有的高独特性值可能预示着严重的误差，如，在 P.H. Weiner 和 J.F. Parcher 所做的研究中，他们利用独特性检验去发现那些可疑的数据点，通过独特性检验，发现丁酸异丁酯显示出独特性，但从化学观点分析，该种溶质同丁酸异丙酯和丁酸异戊酯并没有什么独特之处，因此，认为这种独特性是由所报告的数据中存在的误差所导致的. 另外，他们用 $\lg V_r$ 对碳数目作图发现，除了丁酸异丁酯之外，对其余的溶质均呈线性关系，这也证明了他们关于存在误差的判断的正确性. 他们同时还发现异丁酸异丙酯也具有较高的独特性值，虽然情况并不象丁酸异丁酯那样明显，但当从数据矩阵中将它删除后对所得的子矩阵重新进行因子分析时，发现实验的和预测的检验因子之间有更好的吻合性. 最后决定将这两种酯类从数据矩阵中剔除.

独特性检验还可以用来鉴别相似溶剂当中的聚类. P.H. Weiner 和 D.G. Howery 对一批溶质进行了系统的独特性检验，在一个指定的独特性检验中具有不太高的预测值的溶质可被聚类，这些结果列于表 9.7 中.

仔细观察表 9.7 中醇类的情况，也可以发现独特性检验并非一定聚类化学上相似的指定.

单位检验因子的情况，我们在 3.6 节中已做过简单的说明. 在气 - 液色谱研究中，如果 μ_{ij} 项中的一项对所有的溶质都是恒定的话，则单位向量可被当作一个真正因子来检验. 一般地说，当所有被研究的溶质至少含有一个共同的因子 (如含一个相同的官能团) 时，单位向量就应该被认为是一个因子. 例如，在所研究的烷基 (甲基、乙基、丙基、丁基、戊基、异丙基、异戊基) 丁酸脂和烷基 (乙基、丙基、丁基、戊基、异丁基、异戊基) 异丁酸脂系列中，单位向

量的变换结果是相当好的. 从化学上讲, 这意味着这些溶质的一个特别的自由能贡献同那些与酯链接的官能团的性质无关.

表 9.7 P.H.Weiner 和 D.G.Howery 发表的溶质独特性表

被检验的溶质 (ST)	独特性值	与溶质 ST 成聚类的其它溶质
2.4- 二甲基戊烷	0.34	环已烷
环已烷	0.25	2.4- 二甲基戊烷
苯	0.16	甲苯、苯乙烯
甲苯	0.18	苯、苯乙烯
苯乙烯	0.26	甲苯
甲基碘	0.27	环已烷
丙酮	0.17	丁酮
溴代乙烷	0.07	甲基碘
丁酮	0.19	丙酮
丙醛	0.11	丙酮、丁酮
硝基甲烷	0.35	硝基乙烷
硝基乙烷	0.23	硝基甲烷
吡啶	0.73	–
乙醇	0.19	–
2- 丙醇	0.18	乙醇、正丁醇
正丁醇	0.56	2- 丙醇

9.2.5 典型向量关键集的应用

在气 - 液相色谱中, 由于涉及到大量未知的因素, 故关联和预测保留行为是一件困难的工作. 在 3.5.2 节中, 我们已讨论过, 在组合目标因子分析中应用原始数据的列或行有可能找到典型向量的关键集, P.H. Weiner 等和 R.B. Selzer 等分别用这种方式对气 - 液相色谱数据进行了一系列的研究工作, 结果表明, 保留指数数据可以用小的典型向量集来描述, 每一个可接受的关键集在接近实验误差或是在实验误差范围内复原数据. 在关键集中被描述的分子通常与那些根据化学见解而可能被挑选的分子相似. 对于同一套数据, 许多不同的关键集的组合给出几乎相同的结果, 这一点指明了所采用

的溶质和溶剂的显著冗余. Selzer 等研究了二甲醚等 18 种醚类溶质在 Apiezon L 等 25 种固定相上的保留指数所组成的数据矩阵，通过主成分分析发现，需用 6 个因子才能在 5 个保留指数单位内复原 96% 的数据点，为了找到典型醚类的关键集和典型固定相的关键集，如果去检验每一种可能的组合，那显然是非常费时和费力的. 于是，他们通过下述的方式来试图达到上述目的，在数据复原中，采用典型向量的组合集的任意多个，并将复原的结果情况仔细地记载在一种专门设计的模式表 (见表 9.8) 中. 在较好的数据复原中出现频数最高的那些典型向量可被挑选来构成典型向量的关键集，也就是说，通过审察表中各典型向量的频率趋势，可以估算出每一种向量的相对重要性. 为了便於理解，将其原始数据矩阵中与醚类溶质相对应的行向量的复原情况列于表 9.8 中.

分析表 9.8 中随着因子的数目各典型向量的百分数趋势，我们可以把某一定的数据行与特定的抽象特征向量做试探性的联系. 除二甲醚外，全部其它的醚类都同样地代表第一个特征向量，没有一种醚在此起着主要的作用，这一点与根据醚类的相似性所期望的结果是一致的. 根据在百分比的主要增加情况，当我们从左至右考察表 9.8 中的各个溶质行时，可以见到，二戊基醚所对应的行看起来最近乎与第二个抽象特征向量相当；丁基乙烯醚、烯丙基乙醚或 2-乙基 -1- 己基乙烯醚最近乎与第三个抽象特征向量相当，异丙基丙醚、二甲醚或二异戊基醚与第四个抽象特征向量相当，二丙基或异丁基乙烯醚与第五个抽象特征向量相当，二甲醚或特丁基甲醚则与第六个抽象特征向量相当.

F.V. Warrlent 等曾采集了 101 条紫外 - 可见谱，并对用于选择有代表的波长集的 4 种不同方法比较研究. 结果认为，就总的性能来说关键集因子分析法是最佳的. 他们还继续研究关键集因子分析在色谱中的一种实际应用，在应用色谱对 9 种头孢子菌素溶质进行分离的过程中，他们采用关键集因子分析技术来对用於监控吸光度比率图的波长进行选择.

表 9.8 较好数据复原 (包含矩阵中特定行向量) 中的百分比

序号	溶质行	复原中所用因子数					
		1	2	3	4	5	6
1	甲醚	0	20	0	32	23	100
2	丙基甲醚	6	30	26	33	45	1
3	丁基甲醚	6	30	26	14	3	29
4	特丁基甲醚	6	0	0	15	29	63
5	乙醚	6	20	31	13	13	25
6	丁基乙醚	6	10	0	0	13	37
7	特丁基乙醚	6	0	0	17	0	8
8	丙醚	6	0	0	0	52	28
9	异丙基丙醚	6	0	0	35	16	22
10	二异丙基乙醚	6	0	17	35	55	39
11	特丁基异丙基醚	6	0	13	3	26	28
12	二戊基醚	6	40	17	18	36	40
13	二异戊基醚	6	30	35	63	65	52
14	乙基乙烯醚	6	0	4	24	16	16
15	丁基乙烯醚	6	0	35	21	10	24
16	异丁基乙烯醚	6	0	0	8	52	33
17	2- 乙基 -1 已基乙烯醚	6	20	61	42	36	41
18	烯丙基乙醚	6	0	35	27	10	17
总的组合次数		18	153	816	3060	8568	18564
(舍去的平均的行均误差)		50.0	11.0	6.8	5.0	3.8	3.3
小于舍去的复原次数		17	10	23	78	31	233
最佳组合:							
	平均的行均误差	26.5	9.7	6.1	4.5	3.8	2.8
	所包含的向量	6	3	11,17	2,10	2,8,10	1,4,6,10
			11	18	13,17	13,16	13,16
	最大的误差	266	37	35	24	21	21

在一些 GLC 研究中, 已进行过根据典型向量的关键组合来预测新的保留值数据的工作.

P.H. Weiner 和 D.G. Howery 研究了 L. Rohrschneider 发表的一系列不同的溶质在非极性的角鲨烷和其它溶剂上的保留指数数据矩

9.2.6 目标检验

通过抽象因子分析确定所研究的色谱保留指数数据矩阵的空间维数后，人们自然便要设法去鉴别出那些与抽象特征向量相对应的具有物理或化学意义的参量. 因此，借助目标变换技术去进行溶质和溶剂的参量检验便构成对色谱数据的因子分析研究中的一个最基本的内容.

在应用目标因子分析对气相色谱的开拓性应用研究中， P.H. Weiner 和 D.G. Howery 研究了一个包含有多种多样溶质和溶剂类型的复杂数据集. 他们发现，目标检验分离出许多在化学上被认为是合理的因子. 由于从事气液相色谱 - 目标因子分析的研究人员在发展检验向量方面积累了经验，在色谱理论研究中，目标检验的范围已在增大. 例如， R. B. Selzer 和 D. G. Howery 曾在他们对醚类和各种固定相的相互作用的研究中，检验了 50 多个溶质向量，此外，还检验了每一向量的平方、倒数和对数. 有意义的是，许多被检验参量在不止一个课题的研究中被发现都是因子，有 3 个真实因子 (碳数目、克分子折射和分子量) 至少已在 8 个课题的研究中被判断是溶质因子. 从化学观点来看，这些因子的每一种可能与分散相互作用有联系，当其它因子不变时，洗提与分子量相对应是气液相色谱的一个众所周知的经验规则.

与溶质蒸发有关的参量常能得到好的检验结果. L. Rohrschneider 通过研究指出保留指数与溶质保留作用的自由能变化成正比，因而也就同溶液的焓成正比，溶液的焓可分成溶质的蒸发热和混合的热. 在许多课题的研究中，蒸发焓已被检验证明是溶质因子，根据化学热力学观点，与蒸发焓成比例的参数，如沸点 (K) 和蒸气压的对数都应该是因子. 令人满意的是，这两个参量在一批课题研究中已被确认为溶质因子. 另一类因子是极性相互作用的起因. 偶极矩和偶极矩的平方在某些研究中已被确认为因子，它们都出现在热力学的相互作用项中.

根据目标因子分析的结果可以联想到其它类型的因子. 研究指出，气相非理想性可能与 van der Waals 常数有联系，这些常数在多个气液相色谱问题中都得到好的检验. 目标因子分析能分离出如此小的因子，证实了目标检验的灵敏度. 在几个问题的研究中，代表一种不变的溶质性质的单位向量已通过目标检验，这几个问题所涉及的溶质均具有共同的官能团. 此外，用于具有相似化学性质的溶质的聚类独特性向量往往能成功地被检验，例如，用于芳香族溶质的独特性检验指出苯环在几个问题的研究中是一个因子. 某些色谱工作者不期望成为因子的参数，例如熔点、折射指数、表面张力和粘度等，一般说来检验结果并不满意. 对于溶质，尚有诸如羟基独特性、不饱和独特性、多键标度、羟基位置和单位向量等参量都已在不同的课题中进行过目标检验.

D.G. Howery 等人应用目标因子分析模型对气液相色谱中溶质 - 溶剂之间的相互作用进行了系列的研究并预测保留时间. 他们还对 12 种烃类在多种离子交换树脂上的气 - 固色谱保留指数进行目标因子分析，借以解释溶质 - 吸附剂之间相互作用的程度.

J.K. Strasters 等探索了对于峰鉴别的目标检验的影响. 当用迭代目标因子分析从重叠组分峰中推演出各单个组分的 *UV* 谱时，其结果取决于色谱的分辨率、谱的相似性和各组分的相对浓度. 从观测到的分辨率，观测到的谱相似性和观测到的浓度，他们提出一个定量模型用以判断所推演的 UV 谱的可信度.

G.G.R. Seaton 等通过方差分析技术探讨色谱峰分辨中目标检验的影响，其研究涉及 Ajmalacine, 酞酸二乙酯、长春花碱、酪氨酸、3- 羟基苯甲醛、茶叶碱、甲苯等.

概括地讲，目标因子分析对保留指数的应用研究为溶质 - 溶剂相互作用空间的溶质部分陈述了一个可直观地接受的模型. 目标因子分析已被用来鉴别与分散相互作用、蒸发焓、极性相互作用、气相缺陷和特定的结构参量有关的溶质因子. 总的来说，对於溶质，在目标因子分析结果与化学直觉之间存在的高的相关是一件鼓舞人

技术在色谱理论研究中的强大功能.

9.3 线性自由能关系的研究

对于物理有机化学家来讲，线性自由能关系 (LFER) 是一种很有用的研究工具. 他们可以应用各种 LFER 去预测反应速度和平衡常数、光谱性质和氧化行为、极谱半波电位和偶极矩等. LFER 对於有机反应机理的探讨和搞清楚各种有机分子的化学、物理及生物性质由於结构所诱发的变化都是必不可少的. 在过去的数十年中，LFER 已受到越来越多的重视.

LFER 方法试图用取代基团和反应介质的性质来表达一系列结构上有联系的化合物的反应速度和平衡常数. L.P. Hammett 在 30 年代提出的经验方程式

$$\lg(K_x/K_H) = \rho\sigma_x \tag{9.20}$$

成了 LFER 的起源. 上式中，K_x 和 K_H 分别为已被取代的和未被取代的苯衍生物的速率常数或平衡常数，ρ 是与取代基无关的反应参数，它对于每一反应具有特定性，它反映了诸如介质、温度、压力、反应位置、试剂和支链等可能存在的反应条件所产生的影响，σ_x 则是与反应无关的取代基参数.

Hammett 的方程式原先只用于含有间位和对位取代基的苯化合物，对于邻位取代基的苯化合物，它是不成功的. B. Higman 在试图解释邻位取代基时将抽象因子分析应用到 LFER 中. 他推测 Hammett 方程式中可能包含有由于取代基引起的种种影响而构成的另外的一些项，从而认为一个更合适的表达式可能是

$$\lg K' = \lg K^0 + \sum_i \rho_i\sigma_i, \tag{9.21}$$

式中的加和应包括取代基的全部可能的影响，如反应场、极化、诱

导和位阻等，这个方程式非常适合于因子分析．然而，由于所采用的图解技术的不完备性，他未能得到满意的结果．

E.R. Malinowski 研究了 13 种邻、间和对位取代的苯化合物在 4 种反应介质中的酸度常数，用因子分析对数据进行处理，得出 3 个主要因子．有两个溶质因子被进行目标检验，它们是①数 1，代表 $\lg K^0$ 项的系数乘数；② Hammett σ 常数，从以前只涉及到间位和对位基团的研究中获得．目标检验的结果列于表 9.9 中．

表 9.9　适合于酸度数据的溶质检验向量的目标检验结果

溶质	数 1		Hammett σ 常数		共价半经	
	检验	预测	检验	预测	检验	预测
-H	1	1.004	0.000	0.005	0.00	-0.18
o-F	1	1.04	-	0.690	0.72	0.87
m-F	1	0.93	0.337	0.339	0.00	0.23
p-F	1	1.00	0.062	0.059	0.00	0.07
o-Cl	1	1.07	-	0.823	0.99	0.97
m-Cl	1	1.03	0.373	0.373	0.00	-0.01
p-Cl	1	1.02	0.226	0.219	0.00	-0.03
o-Br	1	1.04	-	0.862	1.14	1.08
m-Br	1	1.00	0.391	0.374	0.00	0.02
p-Br	1	0.99	0.232	0.234	0.00	-0.05
o-I	1	0.90	-	0.920	1.33	1.23
m-I	1	0.98	0.352	0.371	0.00	0.04
p-I	1	0.98	0.276	0.276	0.00	-0.04

从表 9.9 可以看出，对于上面提到的两个因子的检验是成功的．由于缺乏邻位基团的 σ 常数，只好对它们进行自由浮动，用目标检验对这些点做出预测．如表中所列，对邻位上一个取代基的 σ 预测值比相同的取代基在间位和对位上的 σ 预测值要大得多，在当时，这种结果令人感到惊奇．因为一个取代基在邻位上的电子效应一般被假定同其在间位、对位上的电子效应只有少许差异．第三个因子被怀疑是由于取代基在邻位上的位阻现象而造成的，为了解释这一因子的成因，他提出了一个与取代基团的大小有关的检验向量．该

向量由邻位取代基的共价半径和对于间位、对位取代基所设置的零构成. 对这一检验向量进行目标检验的结果也列于表 9.9 中. 对这一目标的检验已给出了完美的相关. 在组合步骤中采用这 3 个因子可以在误差范围内复原酸度常数数据, 因此解释了溶质空间中的全部重要因子. 而且, 从组合目标因子分析所得到的对应于溶剂余因子 ρ 的载荷系数与 H.H. Jaffe 从只包含有间位和对位取代化合物的数据计算所得到的反应介质常数有较好的一致性.

P.H. Weiner 对 19 种带取代基的苯甲酸在 7 种溶剂中的酸度进行了因子分析, 4 个因子被要求来在报道的误差范围内复原数据. 对溶质的独特性检验结果指出, 没有一个溶质在行为上是不规则的. 然而, 在溶剂的情况下, 这种检验表明乙二醇具有独特性, 这种独特性不可能是由氢键引起的, 因为在对于乙二醇的独特性检验中, 其它的醇溶剂显示出低的预测值. 为了解释这 4 个因子的成因, 他提出如下的模型

$$\begin{aligned}\lg(K(i,k)/K^0) = \ & \lg(K(i,\text{gas})/K^0)\cdot 1 + U_E(i,j)V_E(j,k) \\ & + U_W(i,j)V_W(j,k) + U_G(i,j)V_G(j,\text{乙二醇})\end{aligned} \tag{9.22}$$

上式右边的第一项说明在没有溶剂存在时 (即在气相中) 质子转移作用的酸度, 此时, 对气相项取溶剂余因子为 1. 第二项是静电贡献, 第三项是 van der Waals 色散效应, 最后一项说明由乙二醇溶剂产生的独特影响. U 字打头的量与溶质余因子有关, V 字打头的则与溶剂余因子有关. 他还用目标检验来鉴别全部 4 个溶质因子. 由于不要求理论模型, 所以, 溶剂因子中的两个因子 (单位 1 (气相酸度项的溶剂部分) 和对于乙二醇的独特性项) 很容易地被进行目标检验. 对于静电贡献, 他采用了 J.G. Kirkwood 等提出的理论模型

$$(\ln(K(i,a)/K^0))_E = [e\mu_i\cos\theta/2.303RTr^2]_\mu[1/\epsilon_\alpha]_v, \tag{9.23}$$

此模型将静电贡献项表达成溶质项和溶剂项的积函数. 式中 e 是一个电子的电荷, μ_i 为苯甲酸中取代基的键矩, R 为气体常数, T

为溶液温度，r 是取代基偶极矩和酸基团之间的距离，θ 是偶极矩轴和酸的氢之间的角，ϵ_α 表示溶剂介电常数. 上式中的这种积函数形式恰好适合于因子分析. 溶剂的静电因子简单地就是介质的介电常数的倒数. 对于 van der Waals 贡献，他利用了在核磁共振因子分析中 (见节 9.1) 发展出来的一个类似的成功的项. 在该项中，项的溶剂部分被发现是溶剂分子的极性和电子云分布的一个函数. 如表 9.10 所示，对这 4 个溶剂因子所进行的目标检验是成功的.

表 9.10　适合于酸度数据的溶剂因子的目标检验

	单位 1		介电常数的倒数 [a]		van der Waals 效应 [b]		乙二醇的独特性	
溶　剂	检验	预测	预测	检验	检验	预测	检验	预测
甲醇	1	1.02	3.17	2.89	7.44	7.41	0	0.00
乙醇	1	1.00	4.14	4.98	6.93	6.78	0	-0.10
乙二醇	1	0.96	2.66	2.95	7.59	7.59	1	0.93
正丁醇	1	0.97	5.73	5.31	6.78	6.72	0	0.19
戊醇	1	1.03	4.98	4.66	6.85	7.08	0	0.00
二噁烷 - 水 (ϵ= 55)	1	1.00	1.82	1.29	7.84	7.84	0	0.00
二噁烷 - 水 (ϵ= 40)	1	1.00	2.50	3.36	7.72	7.68	0	0.00
二噁烷 - 水 (ϵ= 15)	1	1.00	6.68	6.23	7.46	7.49	0	0.00

a. 检验数据来自 J.H. Elliot 和 Ya.G. Kilpatyick 的工作，$\times 10^{-2}$;

b. 检验数据来自 Ya. G Dorfman 的著作，$\times 10^{-9}$.

对上述 4 个溶剂检验因子组合的同时目标变换产生了酸度实验数据实验误差 3 倍的 RMS 误差. 根据式 9.21，从对溶剂单位 1 项的组合目标因子分析结果所得的余因子应与溶质气相酸度相对应，遗憾的是，在那个时候没有这样的酸度实验数据来作比较. Weiner 通过从数据矩阵中删去乙二醇后重复进行因子分析来验证这些结果的合理性. 如果第四个因子真的只是乙二醇才独有的，则此时的因子空间应该减小为 3，即式 9.21 中的前 3 项便已足够说明问题了. 在独特的溶剂乙二醇被除去后，对式 9.21 的前 3 项，同样得到相同的溶质余因子，它们被列于表 9.11 中.

过下式对实验数据进行标准化

$$S = (I - \bar{I})/\sigma, \tag{9.24}$$

这里，I 是峰强度，$\bar{I}$ 是质量位置的平均强度，σ 是质量位置的标准偏差，获得一个对于平均值的相关阵的修改形式，只保留了那些其方差等于或大于标准化了的质量位置的方差的因子，即只保留了那些其相应的特征值大于 1 的因子. 有 42 个因子被认为是重要的，它们占总方差的 73%. 他们发现，对于羰基类、羟基类、醚类、含氮类、胺类、饱和烃类和含苯类化合物来说，重要的因子数分别为 2, 3, 2, 3, 4, 3 和 6.

R.W. Rozett 等采用了一种更有选择性的研究途径，他们对分子式为 $C_{10}H_{14}$ 的 22 种苯型同分异构体的质谱强度进行了一系列详细的因子分析. 他们推想，由于所有这些烃异构体具有相同的分子量并且都含有一个苯环，故因子空间应该相对简单些，同时，同分异构体的支链结构应提供足够的变化来研究诸如环打开、环膨胀和离子中性络合等已知影响质谱的现象. 在已研究过的各种因子分析的预处理方法中，他们发现，对于质谱来讲，采用绝对强度连同关于原点的协方差矩阵是最好的，用 3 个因子便解释了数据变化的 99.1% 并在 1% 之内复原数据，1% 是在测量质谱线高度中的平均重现性. 对于这样一个复杂的问题，因子空间却小得令人惊奇，异构体 (典型行) 和碎片 (典型列) 的各种集的组合目标因子分析显示出，存在着许多足够地复原数据的异构体和碎片的关键集，由于一些异构体和碎片同样地取决于相同的真实因子，故发现了几个关键集. 反过来，不代表全部 3 个因子的异构体集和碎片集在组合中产生差的复原，这样的研究对将异构体和碎片进行聚类是有帮助的，为检验从质谱因子分析而得到的聚类，他们用三角形图来描述该 3 因子空间. 用 3 个三角形坐标代表 3 个抽象因子，对任一指定的异构体，3 个载荷的平方和被标准化至等于 100，这是对任何在二维平面上画三维图的一个数学要求. 对于主因子解，三角形图显示出异构体的 3 个聚类. 在各个角附近类聚的 16 个、4 个和 2 个异构体分别与第一、第

求得到完整的解. 当处理那些可能具有数以百计的特征向量的大数据矩阵时，这种方法是有价值的. 研究结果表明，有 5 个因子解释了数据的 91% 的变化，它们被认为是可接受的. 采用一种称为直接象限最小的抽象旋转技术来将分子进行聚类，象限最小技术涉及到旋转特征向量以便使载荷余因子的 4 次幂的加和为最小. 通过研究各种化合物对象限最小向量的依赖性，可获得这样的因子解释：这 5 个因子依次与 $C_9H_{11}^+$ 离子、$C_8H_9^+$ 离子、$C_7H_7^+$ 离子、$C_{10}H_{13}^+$ 和 $C_{11}H_{15}^+$ 的存在有关.

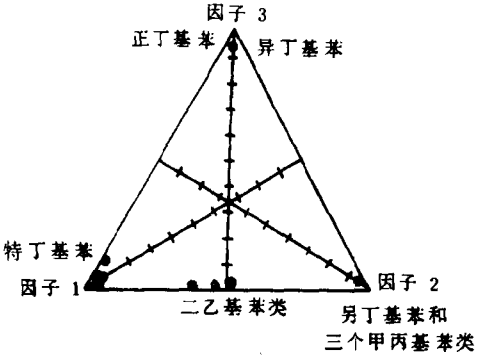


图 9.3　从质谱数据的方差最大旋转得到的载荷平方的三角形图

D.R. Burgard 等研究了寡脱氧核糖核甙酸的质谱以便确定某一特定的核甙是否存在和不同核甙的相对数目以及该化合物中核甙的序列. 原始数据由 32 种选择过的离子的强度组成，这些离子离析自 32 种核甙酸. 这些核甙由 4 种核甙 (腺甙、鸟甙、胸腺嘧啶核甙和胞嘧啶核甙) 的不同的量和不同的连接而成. 通过用同一列中的 32 种离子的强度的总和去除每一个离子强度而归一化每一个与指定的核甙酸有关的每一个数据列，采用关于原点的协方差阵，对归一化后的数据矩阵进行主因子分析，得到说明 99% 变化的 4 个特征向量. 虽

然方差最大旋转指明方差最大因子与 4 种核甙很相似，但它们的载荷不能被用来确定在一个化合物中是否存在一个指定的核甙．为了解决这一问题，采取了以下的步骤．先根据以前所做的质谱研究，挑选 32 种离子中的起始套，一套 8 个被挑选出的离子代表每一个核甙，为了检验某一指定的核甙的存在，通过将与所考虑的核甙有关的子集中的离子强度加和去除每一个离子强度而归一化数据阵中的每一个列，而后再执行主因子分析，从为了确认鸟甙的存在而进行的检验结果所得到的在最前面的两个主因子轴上的因子载荷被画在图 9.4 中． 32 种化合物的载荷形成两个聚类，含鸟甙的化合物同不合鸟甙的化合物被分开，对于其它 3 种核甙也获得了类似的结果．为测定在一个化合物中不同核甙的相对数目，对于每一对核甙被挑选的离子强度的比例被加以考虑，从主因子分析获得的载荷对于在一个化合物中存在的两种核甙的相对量给出了出乎意料的好的估算，遗憾的是，用因子分析来获取序列信息的努力未获得成功．

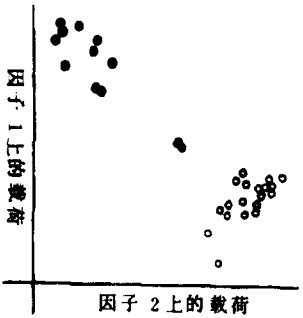


图 9.4 对于经过加和归一的鸟甙离子的因子分析获得的在第一、第二个主因子轴上的载荷的绘图．黑点表示含鸟甙化合物，圆圈则表示不含鸟甙化合物

9.5 在分析化学中的比较

本节中包括了涉及在分析化学中的比较方法的 3 种因子分析应用类型，所有这些方法都利用了已被无意识地融合进因子分析的主

组分特性中的固有统计法.

9.5.1 仪器比较

分析化学家要求有评估分光光度计的运行情况的方法. 基于这种需要, G. Wernimont 检测一组 16 台分光光度计的吸光度曲线, 这些仪器型号相同, 但制造时间却跨越了 20 多年, 在不同的两天内, 在 20 个波长处测量 3 份重铬酸钾溶液的吸光度, 对每台仪器构成了一个 6×20 的数据矩阵, 如果一台仪器运行良好的话, 这一数据矩阵的秩应该为 1, 而如果一台仪器调节校准和读数失当的话, 数据矩阵的秩应该是 2 或更大一些. 对于一个行为理想的组分, 第一个因子便产生于比耳定律; 另外的因子只能是产生于诸如仪器的波长偏移或由化学家引起的大的读数误差等这些误差因素. AFA 简单地可做为例行工作, 定期的应用可使光度计的制作人员确信他的仪器运行正常, 也可以提醒他们什么时候仪器的运行是不令人满意的.

应用关于原点的协方差矩阵, Wernimont 计算了 16 个数据阵中的每一个的残余的标准偏差, 假定秩为 1, 几台仪器显示出比 0.004 吸光单位大的残余标准偏差, 0.004 的吸光单位是从多次单独测量中估算出来的. 在某些情况下, 数据的测量暴露出严重的读数错误, 这易于被纠正, 令人更感兴趣的是, 6 台仪器仍然具有令人不满意的残余偏差, 经仔细机械上的、电方面的和光学方面的调节之后, 所有这些有误差的仪器都产生了可接受的光谱.

9.5.2 方法比较

新的分析方法是分析化学家们的一个追求, 在用一个新方法去取代一个旧方法前, 化学家们必须就系统误差和随机误差以及对该新方法来说可能是特定的干扰物质等方面来评估这一新的方法. 认识到这一问题, R.N. Carey 等应用主因子分析来进行方法比较, PFA 优于稍老些的方法 (如回归分析), 它不需要挑选一个被假定为无误差和无干扰的参照方法, PFA 允许化学家使用不同的混合作为

参照，因而涉及到的是多元统计而不是单变量统计问题.

Carey 及其合作者研究了用以测量葡萄糖的 6 种方法 (列于表 9.12 中)，涉及到被分为 uremic 血清和 nonuremic 血清的共 130 个血清，3 种 PFA 模型被用来评估这些数据. 模型 I 假定有一个系统误差 C_i 和一个随机误差 E_{ik} 存在. 第 i 个方法应用至第 k 个样品所得的葡萄糖的值，Y_{ik} 被表达为

$$Y_{ik} = C_i + E_{ik} + \sum_{j=1}^{m} U_{ij} X_{jk}, \tag{9.25}$$

式中，U_{ij} 是第 j 个物种的响应，X_{jk} 是第 j 个物种的浓度，加和包括了所有对测量有贡献的 m 个物种. 如果无干扰物种存在，m 将等于 1；如果有一个单个的干扰物种，m 将为 2. 根据式 (9.25) 执行 PFA，并对每一种方法计算出在复原的葡萄糖值中的 RMS 误差，对于 uremic 血清，这些误差列于表 9.12 中，标记 E_1 表示 m 是

表 9.12　用以测定 uremic 血清和葡萄糖 6 种方法的 PFA 处理所得的标准误差

	模型 I		模型 II		模型 III	
方法	E_1	E_2	E_1	E_2	E_1	E_2
铁氰化物	10.11	3.47	11.6	3.42	22.48	4.05
Neocuprine	7.06	5.37	6.96	5.41	11.53	5.35
邻 - 甲苯胺	4.21	4.16	4.93	4.24	4.67	4.57
氧化酶 ABTS	6.76	2.64	99.50	2.69	22.56	3.06
氧化酶 MBTH-DMA	3.11	2.15	4.26	2.14	4.14	2.92
己糖激酶	4.77	3.71	5.22	4.16	6.02	4.19

1 的，E_2 表示 m 是 2 的. 模型 II 假定系统误差 C_i 为零，由表中所列的根据模型 II 对 uremic 血清所得的 PFA 结果表明，E_2 的值对模型 I 和模型 II 并没有显著的差别，这就指明了系统误差不存在. 模型 III 假定对于那些使用水的标准校正过的方法，由第一个因子所

引起的响应 U_{i1} 等于 1.00，因而考虑到了取决于所用的方法的校正常数.

对一指定的方法通过比较 E_1 和 E_2，便可迅速得出那些方法对干扰物灵敏的结论. 对一指定的方法，如果 E_2 和 E_1 近似相等，表明无干扰物种存在，如果 E_2 明显的小于 E_1，表明存在一干扰物种，从所有这 3 种模型，可得到相同的结论. 对于 uremic 血清和 nonuremic 血清,铁氰化物和氧化酶 ABTS 法被发现对干扰物质非常敏感，neocuprine 和氧化酶 MBTH-DMA 法对外来物种有中等的敏感性，己糖激酶和邻 - 甲苯胺法相对地不受干扰. 因此，因子分析能揭示出那些葡萄糖测定受到外来物质的干扰或遭到不准确性的损害. 在选择要在例行分析中应用的最合理的分析方法时，这种信息是有价值的帮助.

9.5.3 介质比较

校正分析测量所受介质的影响对于化学家来说已成为一个令人烦恼的问题，在许多情况下，对于有效而廉价地鉴别和补偿介质影响，因子分析可提供有利的帮助.

因子分析用于这一目的的实例可在 J.T. Edward 等的工作中见到. 该工作涉及到羰基化合物在硫酸中的离子化的测量，16 种羰基化合物 (包括醛类、酮类和胺类) 在不同浓度的硫酸中的紫外光谱并不显示出质子化平衡的等交汇点的特征. 研究蒽醌的图 9.5 显示出一个典型的结果，图中顶部的光谱被数字化，所得的数据矩阵被用 PFA 进行处理. 第一个主要特征向量占了总方差的 96%，与质子化影响有关，第二个特征向量占了总方差的 3%，与介质影响有关. 剩余的 1% 变化是由实验误差引起的. 正如图 9.5 下半部的曲线所示，只用第一个主特征向量，由 PFA 重构的曲线显示出清晰的等交汇点. 然后，从这些重构的光谱中，可获得这 16 种羰基化合物的离子化比率和平衡常数.

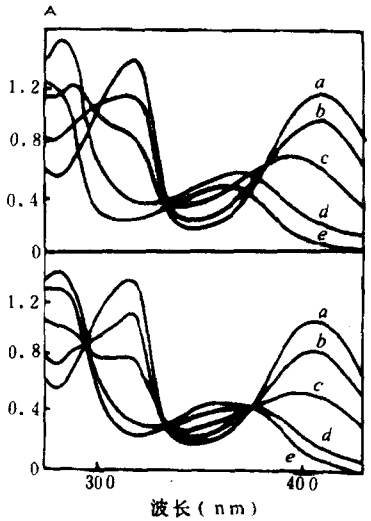


图 9.5 硫酸水溶液中蒽醌的紫外光谱. 酸的浓度分别为 (a) 99.0%, (b) 91.5%, (c) 86.3%, (d) 80.7%, (e)73.0%. 上图是原始数据的谱，下图是根据平均曲线和第一主因子重构的曲线

9.6 其它基础研究

除了前面所述的内容之外，因子分析技术在化学理论的其它一些领域中尚有许多成功应用的例子. 本节将列举部分，望能再给读者一点儿启发.

9.6.1 溶解度与溶液特性

探讨影响气体溶质在液体溶剂中溶解度的因素在物理化学中是一个最基础的问题之一，在化学工程中，气 - 液溶解度的预测也是

一个重要的实际问题. 溶解度数据的因子分析已经被应用于这两种目的.

D.G. Howery 等人用目标因子分析研究了这样一个数据矩阵, 它包含了 8 种非极性溶质 (涉及到氦、氧和乙烷) 和 11 种极性和非极性溶剂 (8 种烷烃和芳烃类、氯代苯、乙醇和二甲亚砜). 从热力学观点来讲, 气相溶质转移至溶液中的自由能已被确认与溶解度的对数而不是溶解度本身成比例, 因此, 他们分析了克分子分数溶解度的对数, 理论上, 自由能被推测是那些与溶质 - 溶剂相互作用有关的项的线性加和, 这些相互作用对数据是有贡献的, 他们对鉴定那些能最好地模型化溶质 - 溶剂相互作用的溶质和溶剂因子感兴趣. 分析结果发现, 4 个因子在 5% (实验误差的上限) 之内复原溶解度矩阵, 误差判据 IND 在 4 个因子处达到最小值更进一步确认了因子的数目, 独特性检验显示出乙烷和甲烷是最独特的溶质、二甲基亚砜和己烷是最独特的溶剂, 采用组合目标因子分析来确定那些能最好地描述数据的典型因子集 (取自数据矩阵的数据向量), 对于两种关键组合集, 复原的 RMS 误差在实验误差范围内. 典型溶质向量的关键集涉及氖、氩、甲烷和乙烷, 而典型溶剂向量的关键集则涉及己烷、戊烷、苯和二甲基亚砜, 最具有独特性的指定都被包括在关键集中, 这是一种在组合因子分析中常被注意到的结果.

为鉴别与溶质 - 溶剂相互作用有关的参数, 他们对许多溶质和溶剂参数进行了目标检验, 其中的几个是根据从前的溶解度理论研究和经验研究而已被提出过的. 过去, 采用溶解度对物质的某些性质所作的图来鉴别因子, 但是, 由于这样的图只有在一个因子的问题中才被期望具有线性. 目标检验为在一个多维的溶解度问题中提供一个更安全的方法, 已被充分地检验过的溶质因子包括分子的质量、蒸发熵、溶解熵、亨利定律常数、极性、硬球直径 (结果见表 9.13) 和 Lennard-Jone 力常数. 被鉴别作为基础因子的溶剂参数是表面张力、蒸汽压对数、刚球直径和一个非极性相互作用项. 著名的希耳德布兰德溶解度参数检验得并不是非常好, 包括检验向量的函数形

式在内的 40 多个溶质因子和近 35 个溶剂向量的检验情况一般或较好.

为了寻找那些描述该相互作用空间的溶质和溶剂部分的最完整模型的参数集，他们进行了基础因子的组合目标因子分析，对于溶质和溶剂，关键组合集给出小于实验误差两倍的 RMS 误差. 亨利定律常数和极性在基础因子的几乎所有组合集中都被描述到，表面张力看起来是一个特别重要的溶剂因子.

D.G. Howery 和他的合作者提出了一个简单的目标因子分析步骤，用来预测新的溶解度数据. 他们的方法包括目标检验溶解度数据的不完整集，因而可预测被自由浮动的分子的溶解度，一个典型的结果列于表 9.13 中，这种预测的精度取决于检验点复盖因子空间

表 9.13　溶质检验向量的目标检验 (对于溶解度数据)

溶质	硬球直径		在环己醇中的溶解度对数	
	检验	预测	检验	预测
He	2.63	2.62	-0.325	-0.327
Ne	2.78	2.79	–	-0.168
Ar	3.40	3.73	-0.724	-0.717
H2	2.87	2.73	0.228	0.238
N2	3.70	3.85	0.424	0.424
O2	3.46	3.36	–	0.712
CH_4	3.70	3.65	1.099	1.100
C_2H_6	4.38	4.35	1.914	1.914

– 表示自由浮动检验点

的程度. 如果相对于独特分子和相对于在典型向量的关键集中被描述的分子的检验点都被包括在向量中，则预测就会更加可靠.

为发展一种实际的预测溶解度数据的方法，C.L. deLigny 等用一种组合抽象因子分析 - 多元回归分析方法来处理 20 种气体在 39 种溶剂中的溶解自由能和溶解熵. 他们提出一种迭代回归步骤来预测数据阵中的空缺数据，然后用抽象因子分析去处理完整的数据矩阵，两个因子解释了自由能和熵数据的主要特性. 通过观察被测量过的值如何从抽象余因子中被满意地加以预测便可检验这种方法的

可行性. 溶解度对数和溶解熵的标准偏差分别为 0.13 和 0.84 卡 · 摩尔 $^{-1}$ 度 $^{-1}$. 对于空缺的数据，溶解度对数的预测值精度被估算为 0.20. 这一方法合理地预测几个点.

为了更好地搞清楚溶液的物理化学特性，W.R. Fawcett 等对几套热力学数据进行了因子分析，这些数据包括诸如溶解焓、溶解自由能、迁移焓和迁移自由能等. 被包括在溶质 - 溶剂数据矩阵中的是极性有机分子和无机电解质. 对于每一种特性，从对于平均值的协方差阵结果所得的两个主要特征向量解释了特性中的 90% 多的方差. 在每一个问题中，两个主因子与溶剂酸性和溶剂碱性的经验估算量有线性联系.

9.6.2 极　　谱

D.G. Howery 对 5 个碱金属离子和 4 个碱土金属离子在 5 种极性溶剂中的极谱半波电位的研究阐明了在一个缺乏理论基础的领域中目标因子分析的应用. 该研究的目的是为离子 - 溶剂相互作用建立一个经验模型，采用 3 个抽象因子，80% 以上的数据在 30 mV 之内被预测. 由于超过 30 mV 的那些实验误差被推测只是对少数几个半波电位而言，故该因子空间看起来可合理地被加以旋转，他应用目标变换来检验几个溶剂和离子的参数，克分子蒸发焓、授体数目

表 9.14　对于极谱半波电位的离子电荷的目标检验

离子	检验值	预测值
Li^+	–	2.03
Na^+	1.0	1.16
K^+	1.0	1.15
Rb^+	1.0	0.92
Cs^+	1.0	0.87
Mg^{2+}	2.0	2.17
Ca^{2+}	2.0	2.06
Sr^{2+}	2.0	1.76
Ba^{2+}	2.0	1.78

和半径校正项作为溶剂向量，其检验情况一般来说是好的. 以这 3 个溶剂因子为基础的模型在组合目标因子分析中给出一个 28 mV 的平均误差值，这对于该问题中的溶剂部分是一个相当好的解. 离子电荷、结晶离子半径的倒数以及它们之间的比率作为离子的向量，其检验情况是好的，目标检验离子电荷的结果列于表 9.14 中. 在检验中，Li^+ 离子的值被审慎地自由浮动，因为已经知道该离子在极性溶剂中有高的过电压. 2.03 的预测值暗示着 Li^+ 离子在这种极谱学研究中的行为象一个二价离子.

9.6.3 稳定常数

D.L. Duewar 及其合作者意识到详细弄清楚影响螯合稳定性的各种因素可能会对改进螯合反应的选择性产生指导作用，因而对 14 种二胺四乙酸配位体和 24 种金属离子的生成常数的对数进行抽象因子分析，因子分析非常适合于络离子的研究，它可给出包括适合于配位体和适合于中心金属离子的因子的解，根据标准自由能变化和平衡常数的对数之间的著名的比例性来进行对数预处理，四个因子在实验误差内解释了数据. 为方便对未经旋转的抽象因子作物理解释，求助于一个由平均稳定常数、金属离子因子和一个配位体因子构成的简化了的三因子模型. 平均稳定常数对应于第一个主要特征向量，离子因子与离子的电荷半径比有联系，但配位体因子却不能被加以鉴别，与所期望的正好相反，配位体的质子缔合常数的对数看起来并不与主因子中的任何一个相对应.

R.L. Reeves 等对含有三配位基染料和 Ni(II) 的溶液的分光光度吸收光谱进行因子分析，借以测定所生成的络合物的生成常数.

S.D. Frans 等采用奇异值分解手段从有机酸混合物的吸收光谱 (随 pH 变化) 获取离解常数和光谱信息.

T. Ozeki 等对 pH 在 5.38—1.79 的 0.03 mM 钼酸钠溶液的紫外光谱进行因子分析，认为共存在 3 种单核钼酸盐型体 (单体、一质子化单体和二质子化单体). 他们并求得两个质子化了的型体的生成常

数分别为 $\lg\beta = 3.773$ 和 7.707 .

9.6.4 键　　能

发展适用于键离解能的理论的和经验的模型是现代化学最重要的目标之一，而因子分析技术特别适合于键离解能的研究. 通过因子分析，一个键离解能可用由某分子离解而生成的两个基团的性质来加以表征. 具体地说，键离解能的一个数据矩阵 ($[E]$) 通过因子分析后可用下式来加以描述

$$[E] = [R_r][R_c], \tag{9.26}$$

式中，$[R_r]$ 和 $[R_c]$ 分别是与行和与列有关的基团性质的矩阵. 服从上式的抽象矩阵可由主成分分析来计算得到，对被分析的数据起主要作用的主要因子数目可用 7.3 节中所讨论过的各种方法来确定. 此外，通过鉴别具有相似行为的基团的组和鉴别具有独特行为的各单个的基团，抽象主因子可被用来关联数据. 最后，代表影响键离解能的基础因子参数可通过目标检验来加以鉴别. 目标检验使键离解能的物理模型得以发展.

D.G. Howery 等对涉及性质不同的 14 种基团 (氢、甲基、乙基、异丙基、特丁基、苯基、苄基、氟、氯、溴、羟基、甲氧基、乙酰基和胺基) 的键离解能数据进行因子分析. 为了研究基团的各种组合的影响，除了处理涉及全部基团的 14×14 对称矩阵外，还研究由该矩阵生成的 5 个对称的子矩阵 (6×6 矩阵涉及从甲基至苄基的 6 种基团； 7×7 矩阵涉及从氢至苄基的 7 种基团； 9×9 矩阵涉及从甲基至溴的 9 种基团； 10×10 矩阵涉及从氢至溴的 10 种基团； 13×13 矩阵涉及从甲基至氨基的 13 种基团), 结果发现，随着问题复杂程度的增加，从因子分析估算得得因子数目也增大 (对于 $6 \times 6, 7 \times 7, 9 \times 9, 10 \times 10$ 和 14×14 的矩阵，根据矩阵复原的 rms 误差和 IND 判据均推知它们的主因子数分别为 2, 2, 2, 3 和 4. 对于 13×13 矩阵则两种判据所得结果分别为 4 和 5). 采用典型向量关键集技术对所研究的 6 个矩阵中的每一个都进行典型向量组合，每次

采用 n 个向量 (n 为各对应矩阵的主因子数)，以寻找最佳模型. 所得结果列於表 9.15 中

表 9.15　典型向量的关键组合集

矩阵	关键基团 [c]	rms 误差 (kcal/mol)[a]	估计误差 (kcal/mol)[b]
6×6	Me, tBu	0.73	2
7×7	H, Et	1.09	2
9×9	Me, F	0.99	2
10×10	H,Me,F	0.94	2
13×13	Me,Ph,F,OH	1.17	3
14×14	Me,Ph,F,OH	1.41	3

a: 对应关键集的 rms 误差； b: 估计实验误差； c: Me: 甲基、 tBu: 特丁基、 H: 氢、 Et: 乙基、 F: 氟、 Ph: 苯基、 OH: 羟基

从表 9.15 中所列 rms 误差值可以看出关键组合集是数据的优秀模型，甲基被包括在 5 个关键组合集中，这表明在总的模型中甲基是碳氢化合物基团的最佳代表. 氟是最重要的卤素基团，而羟基则是 4 个极性基团 (OH, MeO, MeCO 和 NH_2) 中的关键基团. 对于 13×13 以及 14×14 矩阵，其关键组合集由 4 种类型的基团的每一种的一个代表构成，从化学上讲这是合理的.

此外，他们还进行了独特性检验，对于 6 个矩阵， tBu (0.76), H (0.92), F (0.77), H (0.83) 和 F (0.83), F (0.90), F (0.87) 分别呈现出独特性. 对于最后两个矩阵，他们做了目标检验，采用构成分子的基团的近 50 种性质 (如基团的质量、电负性、离子化势、电子亲合性、碳数目等以及含有未配对电子的基团的原子，由相同的基团形成的分子的键离解能等等) 来作为基础因子的候选项.

9.6.5　物质状态

G.R. Rao 等用 FA 和最小二乘曲线拟合去研究十九烷、十九烷 -d_2 及其溶融状态的 FTIR 谱，借以研究有机物分子间的相转变. L.B. Shih 等对不同压力下聚乙烯的喇曼光谱进行因子分析. 他们发现在稍低些的压力下有两个线性无关的贡献 (结晶和熔融物)，而在

最高压力 (3.84 kbar) 下则有 3 个贡献 (结晶、熔融物和中间体).

9.7 在化学其它学科中的应用

9.7.1 生物医学化学

因子分析在生物医学中的种种应用已有报导. 鉴于生物医学数据的固有复杂性和获得可重现的定量测量的困难性, 这些问题较前面一些章节中所讨论过的那些问题要难处理.

在一个早期研究中, M.A. Woodbury 等指出, AFA 可被用来预测生物化学数据. 数据矩阵中的几个点被审慎地略去, 然后被用下面所讲的迭代过程来加以预测, 对空白点设定任意值, 从 AFA 的复原步骤预测空白点的新的估计值, 新的估算值被替代进数据矩阵中, 再次执行复原, 得到又一套估算值, 重复上述过程直至被删去的点得到充分的预测. 对于一个磺胺 - 细胞组织定位矩阵, 预测结果一般不错. 对于一个细菌 - 抗体矩阵, 预测结果相当差, 该方法的一个缺点是在确定用于复原步骤中的因子的合适数目所面临的困难. 事实上, C.G. Swain 及其合作者已经指出, 当数据矩阵中的点丢失时, 迭代因子分析可能导致不合理的结果.

P.H.A. Sneath 用主因子分析研究 20 种氨基酸的化学结构与生物活性之间的关系. 数据矩阵是一个 20×20 的 "相似" 表. 该表是从统计学角度出发被构造来考虑氨基酸的 134 种属性的, 被结合进这一相似表的典型属性包括溶解度、旋光性、色谱保留、各种官能团的存在与否、惯性矩和未共享电子对的数目. 相同的氨基酸代表数据阵的行指定和列指定, 数据矩阵中的每一个元素都是通过一个被设计来测量每一对氨基酸之间的相似性的相关系数来生成的, 由于每一个氨基酸精确地与它本身相似, 所以, 矩阵的对角元素都是 1. 零相似性值指明两种氨基酸没有相似性, 负的相似性值指明相反的特性.

相似性矩阵的主要因子分析产生 4 个具有特征值大于 1 的特

征向量，它们占了总方差的 69%. 通过研究未被旋转的因子载荷，Sneath 将 4 个因子与脂肪族的特性、氢化度、芳香族的特性和巯化度 (涉及羰基和氢硫基形成氢键的能力) 分别建立联系，这 4 个因子被与生物活性相关，然后被用来预测新肽的活性. 虽然准确性不高，但这些预测较那些由偶然猜测所得的预测要好.

对于涉及到生物化合物的模式识别研究，因子分析已成功地被用作一种预处理方法，模式识别因子分析已被用来根据化合物的分子结构将化合物进行治疗学分类. 这种方法要求要有一个总的分子字码编码，它适用于整套化合物中，这种编码必须设计成可以区别原子和基团，从这个总的编码出发可以生成一个结构的特性矩阵，这一矩阵的因子分析可产生特征向量的一个最小集，它可被用于模式识别分类；因而将所要求的轴的数目减至最小，简化了分类和分类的解释. 如果结构的特性确实是相互关联的话，轴的减少是可被期望的. 作为这种技术的一个实例，A. Cammaraton 等用一种称为“超结构”的结构图 (示于图 9.6 中) 来编码了 13 种氨基酸化合物，

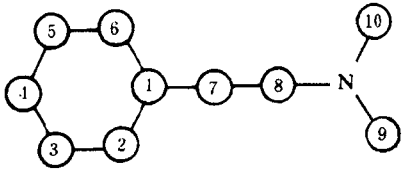


图 9.6　对于 13 种氨基酸化合物的超结构

在这种图中的每一个任意的数字被用来指定在分子 - 特性数据矩阵的一个列，表 9.16 注脚中的字码被用以区别分子的不同特性，用 1 或 0 来指明一个特定官能团的存在与否，用 2 来区别一个芳香族碳和脂肪族碳，对 13 种分子的每一种采用这些特性编码，便产生了分子 - 特性矩阵 (表 9.16). 很容易地验证表 9.16 的第十三行是苄胺的一个编码. 结果所形成的 10×10 相关矩阵的因子分析产生两个

大于 1 的特征值，它们占了总方差的 79%. 行余因子在这两个因子空

表 9.16　亢进剂的分子 - 特性数据矩阵

	特性*									
化合物	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	2	2	2	2	2	2	1	1	0	1
2	1	1	1	1	1	1	1	1	0	1
3	1	1	1	0	1	1	1	1	0	1
4	2	2	2	2	2	2	1	1	1	1
5	1	1	1	0	1	1	0	1	0	0
6	1	1	1	0	0	0	0	1	0	0
7	2	2	2	2	2	2	1	1	0	0
8	1	1	1	0	0	1	0	1	0	0
9	1	1	1	0	0	0	1	1	0	0
10	2	2	2	2	2	2	1	0	0	0
11	2	2	2	2	2	2	0	1	0	0
12	2	2	2	2	2	2	0	0	0	1
13	2	2	2	2	2	2	0	0	0	0

* 根据下面的字码套：

特性	性质	字码	特性	性质	字码	特性	性质	字码
1—6	芳香族原子	2	7	CH_2 存在	1	9,10	CH_3 存在	1
	脂肪族原子	1		CH_2 不存在	0		CH_3 不存在	0
	无原子	0	8	$CHCH_3$	1			
				CH_2	0			

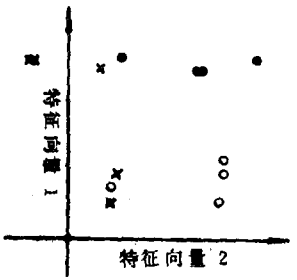


图 9.7　显示弱的 (圆圈)、强的 (黑点) 和未定的 (×) 亢进剂的二维因子空间

间中的作图 (见图 9.7)，图中每一个点与一特定的氨基化合物有联系，这 13 种化合物由亢进剂 (即引起血压升高的药剂) 组成，已知 "强" 和 "弱" 的亢进剂分别用圆点和圆圈来标识，用 × 标识的点的归类事先并不知道，在图上，分子分成明显的两类，如果与已知的生物学响应相匹配，这种根据因子分析的分类看起来是正确的.

在另一个涉及 43 种化合物的分类研究中，G.K. Menon 等应用了以原子和键克分子折射性为基础的编码值，这种编码被设计来考虑在官能团中所存在的“生物等排性”差异，而不只是简单地象前面的例子那样只考虑官能团的存在与否. 生物等排性同那些具有同等的外围电子层的原子、离子和基团有关，这些化合物包括抗组胺、抗抑制剂、抗精神病剂、抗类乙酰胆碱和抗麻痹剂，这一个 43×8 药效基团 - 特性数据阵产生了 4 个大于 1 的特征值，它们占总方差的 79%. 虽然四维因子空间无法画图，但在三维子空间的投影显示出点的聚类，这些聚类与各类医疗活性有联系. 为了进一步简化模式识别研究，他们在一个两步过程中应用因子分析，在第一步中，因子分析被用来将一套 39 种药物分成粗的聚类，以在较小的聚类中的分子为基础的子矩阵然后被单独地进行因子分析，因为小的聚类要求较少一些的判别器，故结果较易解释.

M.L. Weiner 等用目标因子分析法去检验 16 种二苯胺苯酚药物的结构 - 活性相关. 这些药物的特征由在老鼠身上进行的 11 种检验来描述，为满意地复原数据，8 个因子被认为是必须的，因子大小的决定有点儿任意性，因为生物检验中的几个实际上是定性的，对于这些药物，根据独特性检验的结果获得独特性检验值的聚类，对于生物学实验则获得甚至更令人感兴趣的东西，使一种生物试验与其它生物试验的关键集相互联系的目的也会达到.

几个结构性向量和根据独特性检验结果的少数几个向量被进行目标检验. 例如，在药物结构中，一个与氮原子链接的环的存在被显示出是一个因子，经选择过的用于这一检验中的结果列于表 9.17 中，指出了将药物分成含有环链接和没有环链接的这样一种二元分类. 对于最后一个药物的大的预测值看起来好象与该化合物中的双环有关. 有一种预测新药活性的尝试并非太成功，该尝试采用以 8 个典型的生物试验向量的一个关键集为基础的组合 TFA 方法. 当表 9.17 中所示的环链接因子被用来代替关键生物试验向量中的一个时，可观察到对已知药物的预测能力的微小下降.

表 9.17　药物中和氮原子链接的环的单值性检验

药品亚结构 [a]	检验 [b]	预测
$CH_2CH(CH_3)NCH_3$	0.0	0.01
$CH(CH_3)CH_2N(CH(CH_3)_2)_2$	0.0	0.00
$CH_2CH(CH_3)N(CH_3)CH_2C_6H_5$	(0.0)	0.09
$CH(CH_2CH_3)CH_2N(CH_2)_3CH_2$	1.0	0.91
$CH_2CH(CH_3)CH_2N(CH_2)_3CH_2$	(1.0)	1.00
$CH(C_6H_5)CH_2N(CH_2)_3CH_2$	1.0	1.05
$CH_2(CH_2)_3N(CH_2)_3CH_2$	(1.0)	1.0
$CHCH_2NCH_2CH_2CHCH_2CH_2$	–	1.16

a: 完整的分子 $(C_6H_5)_2C(OH)ANRR'\cdot HX$ 的结构性组分 $ANRR'$;

b: 括弧中为已知的点　–: 表示被自由浮动

B.C. Erickson 等采用 Lawton-Sylvestre 自模曲线分辨方法来探讨在临床样品的流动注射分析中背景的影响. R.I. Shrager 等用奇异值分解技术去研究大肠杆菌细胞色素的电位滴定过程，以确定组分从一种状态向另一种状态的转变以及组分的各单个的色谱.

此外，FA 对医药化学的应用包括：用临床和生物检验来研究肝硬化病人的存活时间，服用或不服用一种药物的病人的脑电图，药物对脑细胞膜电位的影响和药片剂的配方等. 涉及到病人对诊断检验和药物对评估性检验的数据矩阵常常可在医疗的和药剂的研究中获得，在这类研究中，因子分析应该能提供有用的分类.

值得注意的是，在生物医学研究中使用这一技术应该特别小心，因为许多类型的医学数据不符合因子分析模型，应经常注意第七章中所讲述过的可因子分析性的判据. D.G. Howery 等研究了这样一个矩阵，它涉及到心脏有问题的病人的血清中的几种组分的浓度，由于没有观察到因子压缩，故因子分析看起来在该问题中并不适用.

级聚类分析被用来进一步澄清因子的性质. 通过一个特殊的步骤对较大的聚类计算因子得分可检测出取样地点之间的相似性，有几个聚类被证明与特定的因子有很强的联系性，例如，一个几乎完全由波士顿市中心地区的各取样点构成的聚类被发现在与汽车尾气有联系的那个因子上有一个大的得分.

P.D. Gaarenstroom 及其合作者用因子分析法研究了在细粒物质中存在的 24 种化学物种的浓度，这些物质是在一年期间内的不同取样时间在阿里桑那州的土松地区的 11 个地点收集得到的，每一个地点的物种 - 时间数据矩阵被单独地进行因子分析，对于不同的取样点，5 至 8 个因子被认为足够复原数据. 通过研究在方差最大旋转解中的载荷，他们将形成污染物的原因归之为 3 个主要来源，即土壤、非地区性气雾和汽车尾气.

E.J. Knudson 等用抽象因子分析法研究了华盛顿 Puget Sound 的雨水，涉及到 22 个取样点的 16 种离子的浓度，3 个因子占了数据总方差的约 70%. 通过应用方差最大旋转，海盐背景、一个广义化的都市源以及一个工业源被解释为是离子的 3 个主要来源.

I.H. Blifford 等就大规模污染、J.T. Peterson 等就气象学参数对空气污染的影响、L.B. Lave 等就空气污染对死亡率的影响和 P. W. Linton 等就都市尘埃的来源都进行了因子分析研究. M.L. Sanchez 等对西班牙某地 9 个取样点的空气中 SO_2 的空间分布进行主成分分析借以鉴别影响 SO_2 浓度的变量. P. Koutrakis 以协方差为基础，对所得的特征向量做斜旋转而达到鉴别和派定俄亥俄州某地空气污染源的目的. M.T. Morandl 等提出了一种改进的因子分析 - 多元回归模型来鉴别和派定在新泽西州 Newark 等地可吸入人体的悬浮颗粒质 (如铅、铁等) 的污染源. D.H. Lowenthal 等应用因子分析探测器模型去模拟城市和地区规模的环境数据集. K. Keiding 提出一种以斜因子分析为根据的用于城市悬浮微粒气体的探测器模型. G.P. Cobb 等人用碳空心管气相色谱法与主因子分析相结合来鉴别大气中的有机物源，分析对象包括调味品混合物、香烟和环境空气.

上述研究阐述了因子分析环境问题的范围，考虑到越来越需要弄明白决定污染物质分布的各种机理，环境数据的因子分析在今后应会更有价值.

9.7.3 其它相关领域

因子分析在其它与化学有关的实践中也有广泛的应用.

在食品和饮料工业中，米的性质同日本米酒风味之间的关系、麦牙的质量与啤酒质量之间的关系、氨基酸成分与土豆的种类之间的关系、氨基酸成分与大麦种类之间的关系已被用因子分析来加以确定.

对地质数据和月亮岩石组成的研究也都利用了因子分析技术. S.L. Bolivar 等用 R 型因子分析描述美国某地区 (19000 km^2) 的 18 套遥感和地球化学数据中的联系，借以进行铀的勘测. 据美国地质勘测部门提供的煤组成数据 (涉及 21 个样品的 78 个成分数据)， B.A. Rossoe 等人采用目标因子分析技术去鉴别和定量分析在煤样品中的各种矿物源，并将所得结果同用 X 射线分析的做了比较.

B.L. Hoesterey 等将从液相色谱分离得到的煤热解焦油的烃油馏分通过两次液相色谱手续再分离成 16 种子馏分，然后采用因子分析与典型相关相结合来处理前 13 种子馏分的低压质谱、红外谱及 H 和 ^{13}C 核磁共振谱，借此来帮助完成对烃油的化学表征工作.

此外，生物系统中的相互作用物种的研究，植物的分类学分类，以及在生态问题中厌氧细菌的菌致分解作用研究等都说明了在各个与化学紧密相联系的学科中因子分析具有广阔的应用前景.

10 多组分同时测定

当一均匀待测混合物体系中的各组分在同一实验条件下只能向人们提供性质非常相似的输出信号时，如需要依赖这些信号来完成对共存组分的同时测定，这时，传统的分析化学观念将促使分析化学工作者首先考虑设法将这些信号源或输出信号彼此进行各种意义上的“分离”，而后再着手去完成测定工作．很自然，如何更巧妙来解决这样的课题也是分析化学家们长期以来一直在思索着的事情．

随着分析测试技术、计算技术和计算工具的迅速发展和谱及，从大量收集到的性质相似的测试信号中直接有效地提炼出分析化学家所关心的有关信号源的化学组成或结构的信息已逐步成为可能．在这方面，适合于多元函数求解的数学的或统计学方面的技术与电子计算机的广泛应用相结合已成为分析化学家手中的一种新的强有力的武器．

近年来，对于性质相似而在测量时给出相互干扰或严重重叠的信号的组分的均匀混合体系的同时测定，化学家们已做了不少努力并已取得了可喜的进展．以分光光度法为例，如稍早些的有解联立方程、线性回归、坐标轮回法等；近期应用得比较成功的则有最小二乘、卡尔曼滤波和因子分析法等计量化学技术．

因子分析法在这个方面具有独到之处，它既能准确地确定混合体系中对输出的信号有贡献的物种的数目，又能鉴别出具体的物种并同时计算出共存的各物种的量．

由于分光光度法应用范围广、采集数据手续简易亦较经济，为帮助从事分析化学工作的读者能较快地学会掌握因子分析法的基本原理并由此而在各自的研究领域中能触类旁通各种因子分析技术，我们打算在这一章中以分光光度法为例，比较通俗浅显地介绍目标

较多的吸光度数据更易于办到. 因此，在讨论中，我们假定 $r < w$. 一般地说，在设计实验时，取 $r > 2n$ 为好.

从数学上讲，吸光度数据矩阵 $[A]$ 完全可以分解成一个行阵 $[R]$ 和一个列阵 $[C]$, 即

$$[A] = [R][C]. \tag{10.4}$$

此时，$[R]$ 为一个 $w \times r$ 阵，$[C]$ 为一个 $r \times r$ 阵.

如果采用行阵 $[R]$ 和列阵 $[C]$ 中的全部数据去进行运算则完全可以复原出原始数据矩阵 $[A]$. 假设对吸光度数据矩阵 $[A]$ 有贡献的吸光物种共有 n 个，那么，取矩阵 $[R]$ 中的 n 列 (构成矩阵 $[R^{\ddagger}]$) 和矩阵 $[C]$ 中的 n 行 (构成矩阵 $[C^{\ddagger}]$) 按式 (10.4) 运算便可得到一个复原阵，假设为 $[A^*]$，那么，$[A^*]$ 就应该在误差范围内与 $[A]$ 相一致. 这时所得到的 $[R^{\ddagger}]$ 和 $[C^{\ddagger}]$ 分别被称为抽象行阵和抽象列阵. 之所以冠以“抽象”一词，那是因为 $[R^{\ddagger}]$ 和 $[C^{\ddagger}]$ 均为纯数学上的抽象结果，它们都并不含有具体的物理或化学上的意义.

在此，重要的问题便是如何从 $[A]$ 来确定 n 的正确值，即如何寻找 $[A]$ 的主因子解. 从线性代数知识可知，寻找 $[A]$ 的主因子解的问题可归结为求解 $[A]$ 的协方差矩阵 $[Z]$ $(= [A]^T[A]$ 是一个 r 阶方阵) 的特征值和特征向量的问题. 求解 $[Z]$ 的特征值和特征向量可采用各种有关的方法，如 JACOBI 法等.

由于实验误差的存在，所求得的 $[Z]$ 的特征值和特征向量数将等于样品溶液的个数 r. 那么，在这 r 个特征向量中，包含有几个是有意义的、即对 $[A]$ 是有贡献的呢 (也就是说，究竟 $n = ?$) ?. 对此，计量化学和统计学工作者已提出了一些很有价值的误差判据，如，RE (真实误差，式 (7.47), IE (嵌入误差式 (7.51)), IND (指示函数，式 (7.56)), ER (特征值比、式 (7.77)), REV (简化特征值，式 (7.66)) 和 MISFIT (3σ 不吻合元素数) 等. 人们可凭借它们的帮助去确定正确的因子数目，即找到正确的 n 值. 到此，产生吸光度矩阵 $[A]$ 的所有混合样品中所存在的对 $[A]$ 有贡献的吸光物种数目的确定的问题已获解决，也就是说，我们已能构成 $[C^{\ddagger}]$ 和 $[R^{\ddagger}]$，以及由 n

个最重要的特征值构成的 $[\lambda^{\ddagger}]$ 阵了. 然而，这时所确定的因子并没有明确的物理或化学意义，仅仅是抽象的数学结果，因此只能说，到此，我们刚完成了抽象因子分析任务.

能否具体地对所存在的 n 个吸光物种具体地加以鉴定呢？能否对每一个样品溶液中的各吸光物种进行定量呢？提出这些问题是非常自然的，也是分析化学家们最为关心的. 目标检验技术的发展，有效地解决了上述问题.

所谓目标，指的是我们认为可能存在于 r 个样品溶液中的某吸光物种，在分光光度法中，可用该物种在相同的实验条件下测得的吸光系数所构成的一个目标向量 $\overline{\overline{R}}_L$ (L 代表第 L 种物种) 来代表目标. 通过下式可求得该物种的目标变换向量 T_L，即

$$T_L = [\lambda^{\ddagger}]^{-1}[R^{\ddagger}]^T\overline{\overline{R}}_L, \tag{10.5}$$

然后，通过下式求出该物种的预测向量 $\overline{R}_L$

$$\overline{R}_L = [R^{\ddagger}]T_L. \tag{10.6}$$

如果，$\overline{R}_L$ 和 $\overline{\overline{R}}_L$ 在实验误差内相匹配，则可判定 $\overline{\overline{R}}_L$ 这个向量所代表的物种就是一个真实因子，即在 r 个样品溶液中确实存在该物种. 这里，又出现一个问题，到底根据什么来判断 $\overline{R}_L$ 和 $\overline{\overline{R}}_L$ 之间的匹配程度呢？ E.R. Malinowski 等提出的 SPOIL(损坏函数判据，式 (7.116)) 和 RELI(可靠性函数，式 (7.114)) 是两种行之有效的目标误差判据. 经过研究，对于运用这两种判据，他们提出这样的经验规则：如果目标的 SPOIL 值在 0—3.0 之间，则该目标可被认为是可接受的；在 3.0—6.0 之间，是尚可接受的；大于 6.0 时，则便是不可接受的了 (即该目标是不存在的)，目标的 RELI 值 $\geq$0.5 时，则被认为是可接受的，否则就是不可接受的. 7.5.4 节所介绍的统计学 F 检验也是一种很好的判据. 通过对怀疑其可能存在于所研究的全部样品溶液中的各个物种逐一进行目标检验，便可完成对所存在的物种的具体鉴定工作. 当然，这时所确定出的 n 种因子就具有明确的物理或化学意义了.

将上面通过检验而鉴定其确实存在的 n 个目标变换向量 (即 T_L, $L = 1,2,\cdots,n$) 组合成目标变换矩阵 $[T]$，然后根据下面的方程式便可求得具有明确的物理或化学意义的吸光系数阵 $[\overline{R}]$ 和浓度阵 $[\overline{C}]$

$$[\overline{R}] = [R^{\ddagger}][T], \tag{10.7}$$

$$[\overline{C}] = [T]^{-1}[C^{\ddagger}], \tag{10.8}$$

$[\overline{R}]$ 的列对应于各物种在不同波长处的吸光系数，$[\overline{C}]$ 的行对应于不同的吸光物种，列则对应于同一样品中各物种的含量.

到此，可以说，因子分析 - 分光光度法用于多组分同时分析的步骤已全部结束.

10.2 氨基酸混合体系的同时测定

为了帮助对因子分析技术不大熟悉的读者更直观地体会该技术的应用，我们在这里介绍一个氨基酸混合体系同时测定的实例.

酪氨酸、色氨酸、苯丙氨酸、胱氨酸、组氨酸和二羟基苯丙氨酸是 6 种在紫外区有吸收的常见氨基酸，它们的紫外吸收谱如图 10.1 所示. 从该图中可以见到，它们的紫外吸收谱重叠严重，在不

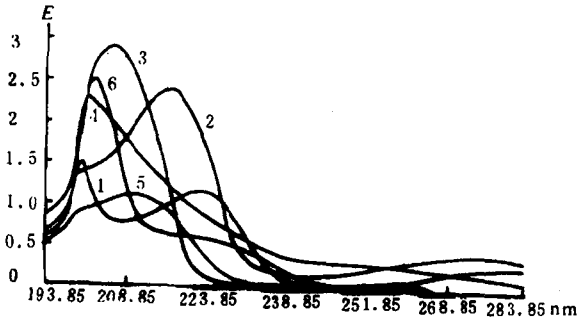


图 10.1 6 种氨基酸的紫外光谱： 1. 酪氨酸 (22.42 ppm); 2. 色氨酸 (12.43 ppm); 3. 苯丙氨酸 (57.34 ppm); 4. 胱氨酸 (197.9 ppm); 5. 组氨酸 (33.72 ppm); 6. 二羟基苯丙氨酸 (16.42 ppm)

1. 抽象因子分析

设从实验量测得到的 12 个含 6 种氨基酸的混合样品溶液在 50 个波长处的吸光度数据构成矩阵 $[A]$

$$[A]=\begin{bmatrix}0.1629 & 0.1603 & 0.1792 & \cdots & 0.1924 & 0.2008 & 0.2337\\ 0.1741 & 0.1713 & 0.1910 & \cdots & 0.2028 & 0.2194 & 0.2472\\ 0.1799 & 0.1764 & 0.1963 & \cdots & 0.2068 & 0.2233 & 0.2518\\ 0.1838 & 0.1812 & 0.2016 & \cdots & 0.2101 & 0.2283 & 0.2547\\ 0.1899 & 0.1876 & 0.2062 & \cdots & 0.2138 & 0.2326 & 0.2562\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ 1.5100 & 1.5031 & 1.5331 & \cdots & 1.6108 & 1.5406 & 1.8570\\ 1.5784 & 1.5719 & 1.6126 & \cdots & 1.6968 & 1.5784 & 1.9031\\ 1.6556 & 1.6459 & 1.7011 & \cdots & 1.7959 & 1.6253 & 1.9508\\ 1.7144 & 1.7122 & 1.7852 & \cdots & 1.8827 & 1.6716 & 1.9830\\ 1.7747 & 1.7670 & 1.8665 & \cdots & 1.9666 & 1.7423 & 2.0269\end{bmatrix}_{50\times 12}$$

计算 $[A]$ 的协方差阵 $[Z](=[A]^T[A])$

$$[Z]=\begin{bmatrix}26.6631 & 26.5449 & 27.3746 & \cdots & 28.5539 & 28.1924 & 32.0062\\ 26.5449 & 26.4275 & 27.2519 & \cdots & 28.4280 & 28.0656 & 31.8657\\ 27.3476 & 27.2519 & 28.1293 & \cdots & 29.3126 & 28.9257 & 32.8426\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ 28.5539 & 28.4280 & 29.3126 & \cdots & 30.6056 & 30.1751 & 34.2524\\ 28.1924 & 28.0656 & 28.9257 & \cdots & 30.1751 & 30.0017 & 33.8489\\ 32.0062 & 31.8657 & 32.8426 & \cdots & 34.2524 & 33.8489 & 38.5207\end{bmatrix}_{12\times 12}$$

通过对 $[Z]$ 进行特征分析 (如用 JACOBI 法) 便可求出 $[Z]$ 的特征值和特征向量. 根据式 (2.79)，用特征向量构成列矩阵 $[C]_{12\times 12}$ 的行，根据式 (2.82)，求出行矩阵 $[R]_{50\times 12}$

$$[C]=\begin{bmatrix}0.259384 & 0.258247 & 0.266264 & \cdots & 0.273834 & 0.311583\\ 0.144848 & 0.135862 & 0.142352 & \cdots & 0.612337 & 0.127631\\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ -0.011901 & 0.333994 & -0.192003 & \cdots & 0.316912 & -0.008700\\ -0.819628 & 0.347574 & 0.013131 & \cdots & -0.082216 & -0.014312\end{bmatrix}_{12\times 12}$$

$$[R]=\begin{bmatrix}0.680474 & 0.012872 & 0.013496 & 0.000854 & -0.000380\\ 0.720143 & 0.016388 & 0.014816 & -0.000164 & -0.000179\\ 0.735247 & 0.018309 & 0.014545 & -0.000091 & -0.000793\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ 6.454497 & -0.217934 & -0.054156 & -0.000726 & -0.001300\\ 6.692914 & -0.251127 & -0.114691 & 0.001456 & 0.000927\\ 6.936088 & -0.259970 & -0.164004 & -0.001299 & -0.000098\end{bmatrix}_{50\times 12}$$

$[Z]$ 的特征值以及根据式 (7.47), (7.51), (7.56), (7.77) 和 (7.66) 计算得到的误差判据列在表 10.1 中.

从表 10.1 中可以看到，在 $n=6$ 之后，IE 值已基本趋于稳定，REV 值则显著变小且大体趋于稳定，$n=6$ 时，IND 值为最小，RE 值为 0.00123，小于 0.0015 吸光单位，ER 值则出现一个折点，综合考虑这些判据，可以断定所研究的这套样品溶液中，因子数目为 6，即存在 6 种氨基酸，这同实验情况完全相符. 因此，用行矩阵 $[R]$ 中的 6 列就可构成抽象行阵 $[R^{\ddagger}]_{50\times 6}$，用列阵 $[C]$ 中的 6 行就可构成抽象列阵 $[C^{\ddagger}]_{6\times 12}$.

到此，抽象因子分析过程全部完成.

表 10.1　特征值和各种误差判据

NF	λ ($\times 10^3$)	RE ($\times 10^3$)	IE ($\times 10^3$)	IND ($\times 10^3$)	ER	REV ($\times 10^3$)
1	395986.7	43.37	12.52	0.358	497.6	660
2	795.75	21.85	8.92	0.219	5.02	1.48
3	158.37	13.37	6.68	0.165	2.23	0.33
4	71.09	4.83	2.79	0.075	10.07	0.165
5	7.06	2.54	1.64	0.052	3.92	0.019
6	1.80	1.23	0.87	0.034	9.19	0.0057
7	0.196	1.02	0.78	0.041	1.59	0.0007
8	0.123	0.82	0.67	0.051	2.16	0.0006
9	0.057	0.72	0.62	0.080	1.53	0.0003
10	0.037	0.64	0.58	0.16	1.71	0.0003
11	0.022	0.61	0.59	0.61	1.15	0.0003
12	0.019					

2. 目标检验与组合变换

这一步骤的目的是要检验某一怀疑的氨基酸是否存在于我们所分析的这套样品溶液中，抽象点讲就是要检验某一怀疑的因子是否真实因子.

为了解决这一问题，在这个例子中，我们以要被检验的氨基酸的不同浓度的标准溶液在相同的实验条件下在相同的波长处所测得的吸光度值为基础，通过线性回归技术求出该种氨基酸在上述波长处的吸光系数来作为它的检验目标.

如要检验酪氨酸是否存在所研究的一套混合样品溶液中，设它在量测原始吸光度数据矩阵 $[A]$ 的那 50 个波长处的吸光系数所构成的检验向量为 $\overline{\overline{R}}_{\text{酪氨酸}} = (r_1 \quad r_2 \quad \cdots \quad r_{50})^T$，如 $r_1 = 5.870 \times 10^{-3}$, $r_2 = 6.341 \times 10^{-3}, \cdots, r_{50} = 2.860 \times 10^{-2}$. 然后按式 (3.14) 计算出它的变换向量 $T_{\text{酪氨酸}}$ =(0.00680 0.06498 -0.00093 0.04624 0.01780 -0.06748)^T，再按式 (3.2) 求出对应的预测向量 $\overline{R}_{\text{酪氨酸}}$ =$(r_1\ r_2 \cdots r_{50})^T$，如 $r_1 = 5.874\times10^{-3}$，$r_2 = 6.312\times10^{-3}$，$\cdots, r_{50} = 2.912\times10^{-2}$.

如果 $\overline{\overline{R}}_{\text{酪氨酸}}$ 是一个真实因子，即酪氨酸确实存在于我们所分析的这套样品溶液中，则 $\overline{\overline{R}}_{\text{酪氨酸}}$ 中的每一个元素应在误差允许范围内与 $\overline{R}_{\text{酪氨酸}}$ 中的对应元素相吻合. 为了对这种吻合的程度做出较定量的判断，可根据式 (7.114) 计算出代表该氨基酸的目标的可靠性函数 RELI=6.117 ，再根据式 (7.116) 计算出该目标的损害函数 SPOIL = 0.7708 . 根据经验规则 (某一目标的 RELI 值等于或大于 0.5 时，该目标可被认为是可接受的，否则就是不可接受的；它的 SPOIL 值在 0.0—3.0 之间，则被认为是可接受的，在 3.00—6.00 之间时是尚可接受的，大于 6.00 时则就被认为是完全不可接受的了). 我们可以做出判断：在这个例子中，代表酪氨酸的目标是可接受的，即酪氨酸存在于我们所分析的这套样品溶液中.

3. 组合和新坐标体系中的列矩阵

按前面所介绍的方法求出全部所存在的氨基酸的对应变换向量

之后，简单地将它们组合在一起就变成一个完整的变换矩阵 $[T]$

$$[T]=[T_{\text{酪氨酸}}\ T_{\text{色氨酸}}\ T_{\text{苯丙氨酸}}\ T_{\text{胱氨酸}}\ T_{\text{组氨酸}}\ T_{\text{二羟基苯丙氨酸}}]$$

根据式 (2.101) 便可求出在新的坐标系下的列矩阵 $[\overline{C}]$

$$[\overline{C}]=[T]^{-1}[C^{\ddagger}]$$

$$[\overline{C}]=\begin{bmatrix} 9.142 & 8.970 & 7.206 & \cdots & 10.990 & 12.759 & 5.579 \\ 2.2640 & 2.193 & 2.284 & \cdots & 2.135 & 2.146 & 5.210 \\ 11.336 & 10.782 & 10.867 & \cdots & 15.235 & 5.277 & 10.173 \\ 54.856 & 54.133 & 78.050 & \cdots & 30.234 & 39.951 & 49.925 \\ 7.080 & 8.176 & 4.057 & \cdots & 7.715 & 10.855 & 7.909 \\ 1.921 & 2.008 & 3.029 & \cdots & 4.299 & 4.311 & 3.186 \end{bmatrix}_{6\times 12}$$

这一矩阵中的每一行代表某一组分在各个混合样品溶液中的含量，如第一行代表酪氨酸在 12 个样品中的含量；而每一列则代表某一样品溶液中各种组分的含量，如第一列则代表第一个混合样品中酪氨酸、色氨酸、苯丙酸氨、胱氨酸、组氨酸和二羟基苯丙氨酸这 6 种组分的含量.

到此，上述 6 种氨基酸多组分混合体系的定性与定量分析结果已全部结束.

10.3 计算一个完整的模拟数值实例

为了帮助初学者检查自己所编写的或是移植得到的计算机程序是否能正确地进行计算，在这里对一个完整的模拟数值实例进行全过程计算并给出各关键步骤的计算结果，以资逐步练习对照.

10.3.1 抽象因子分析

假设有 5 个混合样品溶液，在 10 个波长处量测吸光值，构成原始吸光数据矩阵 $[A]$

$$[A]_{10\times5}=\begin{bmatrix}0.3400 & 0.2260 & 0.4604 & 0.4881 & 0.8982\\ 0.4498 & 0.2999 & 0.6093 & 0.6558 & 1.1890\\ 0.5091 & 0.3934 & 0.7980 & 0.8672 & 1.5573\\ 0.7256 & 0.4837 & 0.9747 & 1.0981 & 1.9026\\ 0.8666 & 0.5777 & 1.1582 & 1.3415 & 2.2612\\ 0.9411 & 0.6267 & 1.2464 & 1.5057 & 2.4342\\ 0.9354 & 0.6236 & 1.2198 & 1.6000 & 2.3839\\ 0.8185 & 0.5457 & 1.0721 & 1.3816 & 2.0930\\ 0.6499 & 0.4332 & 0.8195 & 1.2511 & 1.6039\\ 0.3846 & 0.2564 & 0.4687 & 0.8225 & 0.9186\end{bmatrix}_{10\times5}$$

计算 $[A]$ 的协方差阵 $[Z]=[A]^T[A]$.

$$[Z]=\begin{bmatrix}4.943822 & 3.295169 & 6.516335 & 8.105421 & 12.72786\\ 3.295169 & 2.196304 & 4.343281 & 5.402464 & 8.483401\\ 6.516335 & 4.343281 & 8.59357 & 10.66075 & 16.78482\\ 8.105421 & 5.402464 & 10.66075 & 13.40351 & 20.82459\\ 12.72786 & 8.483401 & 16.78482 & 20.82459 & 32.78386\end{bmatrix}_{5\times5}$$

通过对 $[Z]$ 进行特征分析 (如用 JACOBI 法) 便可求出 $[Z]$ 的特征值和特征向量. 根据式 (2.79) ，用特征向量构成列矩阵 $[C]_{5\times5}$ 的行，再根据式 (2.82) 求出行矩阵 $[R]_{10\times5}$

$$[C]=\begin{bmatrix}0.282863 & 0.188534 & 0.372791 & 0.463973 & 0.728148\\ 0.030870 & 0.020436 & 0.225092 & -0.875739 & 0.425494\\ 0.668834 & 0.218775 & 0.488428 & -0.092422 & -0.507637\\ -0.64074 & -0.056185 & 0.741831 & 0.084187 & -0.169985\\ -0.247283 & 0.955520 & -0.146580 & -0.046707 & -0.046539\end{bmatrix}_{5\times5}$$

$$[R]=\begin{bmatrix}1.19062 & 0.073446 & -0.000021 & 0.000041 & 0.000035\\ 1.580956 & 0.088765 & -0.000140 & 0.000040 & 0.000057\\ 2.074876 & 0.109060 & -0.000181 & 0.000066 & 0.000031\\ 2.554662 & 0.099577 & -0.000121 & -0.000003 & 0.000052\\ 3.05472 & 0.086582 & -0.000158 & 0.000032 & 0.000050\\ 3.320066 & 0.039550 & 0.000472 & -0.000611 & -0.000201\\ 3.315079 & -0.070661 & -0.000192 & -0.000028 & 0.000082\\ 2.899115 & -0.041622 & 0.000295 & 0.000745 & -0.000057\\ 2.319361 & -0.199809 & -0.000113 & -0.000139 & 0.000022\\ 1.382351 & -0.206824 & -0.000079 & -0.000043 & 0.000022\end{bmatrix}_{10\times5}$$

$[Z]$ 的特征值以及根据式 (7.47), (7.51), (7.56), (7.77) 和式 (7.66) 计算得到的误差判据列在表 10.2 中.

表 10.2 特征值以及相应的各种误差判据

NF	λ	RE $\times 10^4$	IE $\times 10^4$	IND $\times 10^4$	ER	REV $\times 10^3$
1	61.7875	577.87	258.4	36.12	426.6	1235.7
2	0.133569	3.85	2.44	0.43	47003.3	3.71
3	2.84×10^{-6}	2.84	2.20	0.71	1.83	0.00012
4	1.56×10^{-6}	0.75	0.67	0.75	27.38	0.00011
5	5.69×10^{-8}					

从表 10.2 可以见到，$n=2$ 之后，IE 值已基本趋于稳定，REV 值则显著变小且趋于稳定，$n=2$ 时，IND 值为最小，RE 值为 0.00039，大大小于 0.0015 吸光单位，ER 值则出现一个折点，对这些判据的综合考虑，可以断定所研究的矩阵的因子数为 2. 因此，用行阵 $[R]$ 中的 2 个列就可构成抽象行阵 $[R^{\ddagger}]_{10\times2}$，用列阵 $[C]$ 中的 2 个行就可构成抽象列阵 $[C^{\ddagger}]_{2\times5}$，用 $[R^{\ddagger}]$ 和 $[C^{\ddagger}]$ 就可得到一个复原矩阵 $[A^*]$，它在误差范围内同矩阵 $[A]$ 相等.

到此，抽象因子分析过程全部结束.

10.3.2 目标检验与组合变换

以某一组分在讨论中的 10 个波长处的吸光系数代表该组分的检验目标. 检验完毕并确定所存在的 2 个因子后，再进行组合变换，便可求出 5 个混合样品中这 2 个因子的含量. 假设这 2 个因子的目标向量分别为 $\overline{\overline{R}}_1$ 和 $\overline{\overline{R}}_2$. 按式 (10.5) 和 (10.6) 可计算出它们各自的变换向量 T_1, T_2 和预测向量 $\overline{R}_1$ 和 $\overline{R}_2$.

$$T_1=\begin{bmatrix}4.554899\\18.790710\end{bmatrix},\quad T_2=\begin{bmatrix}4.845826\\-18.004100\end{bmatrix}.$$

$\overline{\overline{R}}_1$, $\overline{R}_1$, $\overline{\overline{R}}_2$ 和 $\overline{R}_2$ 的数据列于表 10.3 中.

表 10.3 目标检验的结果

$\overline{\overline{R}}_1$	$\overline{R}_1$	$\overline{\overline{R}}_2$	$\overline{R}_2$
6.8028	6.8032	4.4481	4.4472
8.8704	8.8691	6.0622	6.0629
11.5008	11.5002	8.0906	8.0910
13.5072	13.5073	10.5876	10.5867
15.5417	15.5409	13.2442	13.2438
15.8648	15.8657	15.3755	15.3764
13.776	13.7721	17.3335	17.3365
12.4177	12.4321	14.8008	14.7980
6.8098	6.8099	14.8375	14.8366
2.4107	2.4101	10.4221	10.4223

将检验向量 T_1 和 T_2 组合变换矩阵 $[T]$

$$[T] = \begin{bmatrix} 4.554898 & 4.845826 \\ 18.790710 & -18.004100 \end{bmatrix},$$

然后，再根据式 (10.8) 便可求得各样品中各组分的浓度矩阵

$$[\overline{C}] = \begin{bmatrix} 0.0303 & 0.0202 & 0.0451 & 0.0237 & 0.0877 \\ 0.0299 & 0.0199 & 0.0345 & 0.0734 & 0.0679 \end{bmatrix}.$$

到此，计算过程已全部结束.

事实上，通过复算便可发现用 $\overline{\overline{R}}_1$ 和 $\overline{\overline{R}}_2$ 可以构成一个 10×2 的行矩阵 $[\overline{R}]$，将 $[\overline{R}]$ 与 $[\overline{C}]$ 相乘就会得到本节中得矩阵 $[A]$.

10.4 目标因子分析法的一种改进

前面介绍过的目标因子分析法是以在一定的量测条件下各吸光物种均符合朗伯 - 比耳定律且各物种之间线性无关这些假设为前提的. 但是在分析化学实践中，情况有时并非如此理想，尤其是混合样品中共存的吸光物种数较大时，由于共存交互作用，各物种的光学行为偏离朗伯 - 比耳定律，且样品的吸收曲线也会发生变化，表

距因子之中，因此，如再次判断因子数未发生变化，说明存在非零截距因子.

共存的物种之间可能会发生相互作用，致使吸收曲线发生变化，采用更接近实际分析情况的模拟纯谱来替代实测各单独物种所得的纯谱来进行计算，结果应该更准确. 为获得各吸光物种的模拟纯谱以及所研究体系在实验条件下的非零截距因子，可用已知的各吸光物种来配制成标准混合样品，配制的原则是使其具有充分的代表性. 标准混合样品中各物种的浓度之间应成各种比例，浓度的波动范围应同所研究的体系相类似，而且各物种间的浓度向量应该是线性无关的.

对量测标准混合样品所得的数据矩阵的转置进行因子分析，然后以各物种的浓度向量及一个单位向量作为目标，按 10.1 节所介绍的方法便可求出各物种的模拟纯谱以及该混合样品体系的非零截距因子. 然后，以这些模拟纯谱以及非零截距因子 (为运算方便最好都应先进行归一化处理) 作为目标，去对待测的混合样品进行目标因子分析.

实践证明，对于某些混合体系，这种改进的目标因子分析法比通常的目标因子分析法具有更高的准确性.

10.4.2 实际应用

应用这种改进的目标因子分析法测定去痛片的组成及含量取得了准确的结果. 去痛片的基本成分是咖啡因、非那西丁、氨基比林和苯巴比妥. 由于苯巴比妥在酸性溶液中紫外区没有吸收，故在测定中采用硼酸氯化钾缓冲溶液 (pH= 9.1) 作为介质，选择的波长范围为 220—272nm，量测间隔为 4nm. 在这些实验条件下，4 种组分都有较大的吸收.

采用上述 4 种组分的纯标准溶液，按正交设计原则配制成一套标准混合样品. 在选择的波长范围内量测吸光度值构成原始吸光数据矩阵，将其转置后，进行因子分析. 取 4 种组分的浓度向量及一

个单位向量为目标，求出该套样品的非零截距因子及各纯组分的模拟纯谱．取因子数为 5，以已求得的模拟纯谱和非零截距因子为目标，用改进的目标因子分析法去测定实际样品并同通常的目标因子分析方法作了比较，部分结果列于表 10.4 中．

从表 10.4 可以看出，对于象去痛片这类的混合样品，用通常的目标因子分析法所测得的结果误差很大，而用本节所介绍的改进目标因子分析法，则结果是比较满意的．

表 10.4 改进的 (MTFA) 与通常的 (TFA) 目标因子法测定结果比较 ($\mu g/ml$)

	咖啡因			非那西丁		
	理论值	MTFA	TFA	理论值	MTFA	TFA
1	3.97	3.76	5.52	12.15	12.12	13.45
2	3.97	3.73	5.99	12.15	12.07	13.86
3	3.75	3.72	5.51	11.25	11.50	12.88
4	3.75	3.63	5.35	11.25	11.39	12.70
5	3.75	3.78	5.64	11.25	11.53	12.97
6	3.75	3.67	5.53	11.25	11.45	12.90
7	3.75	3.61	5.43	11.25	11.38	12.79
	氨基比林			苯巴比妥		
	理论值	MTFA	TFA	理论值	MTFA	TFA
1	12.03	12.84	8.67	1.32	1.42	0.90
2	12.03	12.85	7.78	2.40	2.52	1.85
3	11.25	11.74	7.41	1.13	1.18	0.60
4	11.25	11.86	7.72	1.13	1.16	0.62
5	11.25	11.67	7.17	1.13	1.16	0.54
6	11.25	11.88	7.36	1.13	1.17	0.54
7	11.25	11.99	7.58	1.13	1.18	0.58

参 考 文 献

[1] Malinowski E.R., Howery D.G.. *Factor Analysis in Chemistry*, Wiley, New York: 1980

[2] Malinowski E.R.. *Factor Analysis in Chemistry*, 2nd ed. Wiley, New York: 1991

[3] Ramos L.S., Beebe K.R., Carey W.P., Eugenio Aanchez M., Erickson B.C., Wilson B.E., Wangen L.E., Kowalski B.R.. *Anal. Chem.*, **58**, 294R–315R(1986)

[4] Brown S.D., Barker T.Q., Larivee B.J., Monfre, S.L.,Wilk H.R.. *Anal. Chem.*, **60**, 252R–273R(1988)

[5] Brown S.D.. *Anal. Chem.*, **62**, 84R–101R(1990)

[6] 俞汝勤，化学计量学导论，长沙，湖南教育出版社，1991

[7] 张尧庭，方开泰，多元统计分析引论，北京，科学出版社，1983

[8] 何锡文，任洪吉，史慧明，分析化学，**15**(4),372–381(1987)

[9] Massart D.L., Vandeginste B.G.M., Deming S.N., Michotte, Y., Kaufman L.. *Chemometrics: a textbook*, Elsevier, New York: 1988

[10] 殷龙彪，分析化学，*16*(8), 761–766(1988)

[11] 司圣柱，潘忠孝，张懋森，光谱学与光谱分析，**9**(3), 57–61(1989)

[12] Kankare J.J.. *Anal. Chem.*, **42**, 1322(1970)

[13] Wallace R.M.. *J. Phys. Chem.*, **64**, 899(1960)

[14] Jochum C., Kowalski B.R.. *Anal. Chim. Acta*, **133**(4), 583(1981)

[15] 许志宏，刘学文，郑修贵，夏永年，查金荣，TQ-16 机 FORTRAN 语言常用算法程序集，北京，化学工业出版社，1982

[16] 忻新泉，计算机在化学中的应用，南京，南京大学出版社，1986

[17] [美] Johnson K.J. 著，汤定华，于建国，郑世钧，郝金库译，电子计算机在化学中的应用，北京，北京师范大学出版社，1988

[18] Malinowski E.R.. *Anal. Chim. Acta*, **134**, 129–137(1982)

[19] McCue M., Malinowski E.R.. *Anal. Chim. Acta*, **133**, 125–136 (1981)

[20] McCue M., Malinowski E.R.. *Apllied Spectroscopy*, **37**(5), 463–469 (1983)
[21] Brayden T.H., Poropatic P.A., Watanabe J.L.. *Anal. Chem.*, **60**, 1154–1158(1988)
[22] Lorber A., Kowalski B.R.. *Anal. Chem.*, **61**, 1168–69(1989)
[23] Rummel R.J.. *Applied Factor Analysis*, Northwestern University Press, Evanston: I11., 1970
[24] Ho C.N., Christian G.D., Davidson E.R.. *Anal. Chem.*, **50**(8), 1108–1113(1978)
[25] Ho C.N., Christian G.D., Davidson E.R.. *Anal. Chem.*, **52**(7), 1071–1079(1980)
[26] Ho C.N., Christian G.D., Davidson E.R.. *Anal. Chem.*, **53**(1), 92–98(1981)
[27] Gianelli M.L., Burns D.H., Callis J.B., Christian G.D., Andersen N.H.. *Anal. Chem.*, **55**(12), 1858–1862(1983)
[28] Lorber A.. *Anal. Chim. Acta*, **164**, 293–297(1984)
[29] Rossi T.M., Warner I.M.. *Anal. Chem.*, **58**(4), 810–815(1986)
[30] Sanchez E., Kowalski B.R.. *Anal. Chem.*, **58**(2), 496–499(1986)
[31] McCue M., Malinowski E.R.. *J. Chromatogr. Sci.*, **21**, 229–234(1983)
[32] Lorber A.. *Anal. Chem.*, **57**(12), 2395–2397(1985)
[33] Shrager R.I.. *Chemometrics Intell. Lab. Syst.*, **1**, 59–70(1986)
[34] Wilson B.E., Lindberg W., Kowalski B.R.. *J. Am. Chem. Soc.*, **111**, 3797(1989)
[35] Wilson B.E., Kowalski B.R.. *Anal. Chem.*, **61**, 2277–84(1989)
[36] Gampp H., Maeder M., Meyer C.J., Zuberbuhler A.D.. *Talanta*, **32**(2), 95–101(1985)
[37] Gampp H., Maeder M., Meyer C.J., Zuberbuhler A.D.. *Talanta*, **32**(4), 257–264(1985)
[38] Gampp H., Maeder M., Meyer C.J., Zuberbuhler A.D.. *Talanta*, **32**(12), 1133–1139(1985)
[39] Gampp H., Maeder M., Meyer C.J., Zuberbuhler A.D.. *Talanta*, **33**(12), 943–951(1986)
[40] Malinowski E.R. in Meuzelaar H.L.C. *et al.*. Ed.. *Computer-Enhanced Analytical Spectroscopy*, Plenum Press, New York: 1987

[67] Gillette P.C., Lando J.B., Koenig J.L.. *Anal. Chem.*, **55**, 630(1983)
[68] Meister A.. *Anal. Chim. Acta*, **161**, 149–161(1984)
[69] 李通化. 分析化学, **15**, 887(1987)
[70] Gemperline P.J., Boyette S.E., Tyndall K.. *Appl. Spectrosc.*, **41**, 454(1987)
[71] Nakanishi K., Maya K., Howery D.G.. *Bull. Chem. Soc. Jap.*, **60**, 2677(1987)
[72] 何锡文, 任洪吉, 史慧明, 郭登峰, 侯晓清. 分析化学, **15**, 495(1987)
[73] 潘忠孝, 夏四清, 张懋森, 刘信安, 石乐明, 李志良. 分析化学, **19**, 826–830(1991)
[74] Pan zhongxiao, Xia Siqing, Si Shengzhu, Zhang Maosen, Shi Leming, Liu Xinan. *J. Chemometrics*, **4**, 323–330(1990)
[75] 夏四清, 潘忠孝, 张懋森, 刘信安, 石乐明, 李志良. 中国科学技术大学学报, **20**(4), 428–436(1990)
[76] Rusmassen G.T., Isenbour T.L., Lowry S.R., Ritter G.L.. *Anal. Chim. Acta*, **103**, 213(1978)
[77] Antoon M.K., Desposito L., Koenig J.L.. *Appl. Spectrosc.*, **33**, 351 (1979)
[78] Bulmer J.T., Shurvell H.F.. *J. Phys. Chem.*, **77**, 256(1973)
[79] Bulmer J.T., Shurvell H.F.. *Can. J. Chem.*, **53**, 1251(1975)
[80] Korppi-Tommola J., Shurvell H.F.. *Can. J. Chem.*, **56**, 2959(1978)
[81] Lawton W.H., Sylvestye E.A.. *Technometrics*, **13**, 617(1971)
[82] Ohta M.. *Anal. Chem.*, **45**, 553(1973)
[83] Yamako k., Takatsuki M.. *Bull. Chem. Soc. Jap.*, **51**, 3182(1978)
[84] Lin C.H., Lin S.C.. *J. Chin. Chem. Soc.*, **25**, 167(1978)
[85] 李科, 陶元, 王宗明, 中国化学会第一届全国计算分析会议交流材料, 合肥, 1985
[86] Gillette P.C., Koenig J.L.. *Appl. Spectrosc.*, **36**, 535(1982)
[87] Fridericks H.B., Yu J.P.. *Appl. Spectrosc.*, **41**, 454(1987)
[88] Fredericks P.M., Lee J.B., Osborn P.R., Swinkels D.J.. *Appl. Spectrosc.*, **39**, 303–310; 311–316(1985)
[89] Haaland D.M., Thomas E.V.. *Anal. Chem.*, **60**, 1202–1208(1988)

[111] Bentley G.E.. Hamilton V.T., Peterson E.J., Wangen L.E.. *Appl. Spectrosc.*, **40**, 949–953(1986)
[112] Lorber A., Eldan M., Golbart Z.. *Anal. Chem.*, **57**, 851–857(1985)
[113] MacNaughtan D. Jr., Rogers L.B., Wernimont G.. *Anal. Chem.*, **44**, 1421(1972)
[114] Davis J.E., Shepard A., Stanford N., Rogers L.B.. *Anal. Chem.*, **46**, 821(1974)
[115] Frans S.D., Mcconnell H.L., Harris J.M.. *Anal. Chem.*, **57**, 1552–1559(1985)
[116] Eide M.O., Kvalheim O.M., Telnaes N.. *Anal. Chim. Acta*, **191**, 433–437(1986)
[117] Gemperline P.J.. *Anal. Chem.*, **58**, 2656–2663(1986)
[118] Ritler G.L., Lowry S.R., Isenhour T.L., Wilkins C.L.. *Anal. Chem.*, **48**, 591(1976)
[119] Windig W., Jakab E., Richaras J.M., Meuzelaar H.L.C.. *Anal. Chem.*, **59**, 317(1987)
[120] Windig W., Liebman S., Wasserman M.B., Snyder A.P.. *Anal. Chem.*, **60**, 1503(1988)
[121] Liu X.D., Michiels F., Van Espen P., Adams F.. *Mikrochim. Acta*, **3**, 49–70(1986)
[122] Tsao R., Voorhees K.J.. *Anal. Chem.*, **56**, 1339–1343(1984)
[123] Koemig S., Hoogerbrugge R., Van Witzenburg W.R., Kistemaker P.G.. *Int. J. Mass spectrom. Ion Processes*, **89**, 111/24(1989)
[124] Malinowski E.R., McCue M.. *Anal. Chem.*, **49**, 284(1977)
[125] Vallis L.V, Macfie H.J., Gutteridge C.S.. *J. Anal. Appl. Pyrol.*, **5**, 333(1983)
[126] Answorth S.. *J. Phys. Chem.*, **65**, 1968(1961)
[127] Answorth S.. *J. Phys. Chem.*, **67**, 1613(1963)
[128] Katakis D.. *Anal. Chem.*, **37**, 876(1965)
[129] Cochran R.N., Horne F.H.. *Anal. Chem.*, **49**, 846(1977)
[130] Crociani B., Horne F.H.. *Chem. Soc. Dalton Trans.*, **12**, 2303(1982)
[131] Haldna V., Horn F.H.. *Comput. Chem.* **8**, 201(1984)
[132] Giibert R.A., Liewellyn J.A., Swarte Jr. W.E.. *Appl. Spectrosc.*, **39**, 316(1985)

[133] Halaka F.G., Babcock G.T., Dye J.L.. *Biophys. J.*, **42**(2), 209–219(1985)
[134] McMullen D.W., Jaskunas S.R., Tinoco Jr I.. *Biopolymers*, **5**, 589 (1967)
[135] Starks T.H., Fang J.H., Zevin L.S.. *Math. Geol.*, **16**, 351–367(1984)
[136] Koenig M.F., Grant G.T.. *J. Electron Spectrosc Relat. Phenom.*, **41**, 145(1986)
[137] Solomon J.S.. *Thin Solid Films*, **154**, 11–20(1987)
[138] Gaarenstroom S.W.. *J. Var. Sci. Technol.*, **16**(2), 600(1979)
[139] Gaarenstroom S.W.. *Appl. Surf. Sci. Technol.*, **7**, 7–18(1981)
[140] Gaarenstroom S.W.. *J. Var. Sci. Technol.*, **20**, 458–461(1982)
[141] Gaarenstroom S.W.. *Appl. Surf. Sci.*, **26**, 561–574(1986)
[142] Wandass J.H., Turner N.H.. *J. Vac. Sci. Technol.*, **A, 6**, 1027-1031(1986)
[143] Solomon J.S.. *Surf. Interface Anal.*, **10**, 75–86; 216–218(1987)
[144] Solomon J.S., Smith S.R.. *Anal. Chem.*, **58**, 51–57(1986)
[145] Solomon J.S.. *SIA, Surf. Interface Anal.*, **10**, 216–218(1987)
[146] Hoffman S., Steffen J.. *SIA, Surf. Interface Anal.*, **14**, 59/65(1989)
[147] Kargancin M.E., Kowalski B.R.. *Anal. Chem.*, **58**, 2300–2306(1986)
[148] Homer J.. *Appl. Spectrosc.*, **9**, 132(1975)
[149] Weiner P.H., Malinowski E.R., Levinstone A.R.. *J. Phys. Chem.*, **74**, 4537(1970)
[150] Buckingham A.D., Schaefer T., Schneider W.G.. *J. Chem. Phys.*, **32**, 1227(1960)
[151] Malinowski E.R., Pierpaoli A.R.. *J. Magn. Reson.*, **1**, 509(1969)
[152] Schug J.C.. *J. Phys. Chem.*, **70**, 1816(1970)
[153] Weiner P.H., Malinowski E.R.. *J. Phys. Chem.*, **75**, 1207(1971)
[154] Weiner P.H., Malinowski E.R.. *J. Phys. Chem.*, **75**, 3160(1971)
[155] Linder B.. *J. Chem. Phys.*, **33**, 668(1960)
[156] Howard B.B., Linder B., Emerson M.T.. *J. Chem. Phys.*, **36**, 485(1961)
[157] Rummens F.H.A., Raynes W.T., Bernstein H.J.. *J. Phys. Chem.*, **72**, 211(1968)

[158] Bacon M.R., Maciel G.. *J. Am. Chem. Soc.*, **95**, 2413(1973)
[159] Abraham R.J., Wileman D.F., Bedford G.R.. *J. Chem. Soc. Perkin Trans.*, **II**, 1027(1973)
[160] Azzaro M., Geribaldi S., Videau B., Chastrette M.. *Org. Magn. Reson.*, **22**, 11/15(1984)
[161] Weiberg K.B., Pratt W.E., Bailey W.F.. *Tetrahedron Let.*, **49**, 4861–4865(1978)
[162] Rohrschneider L.. *J. Chromatogr.*, **22**, 6(1966)
[163] Funke P.T., Malinowski E.R., Martire D.E., Pollara L.Z.. *Sep. Sci.*, **1**, 661(1966)
[164] Wold S., Andersson K.. *J. Chromatogr.*, **80**, 43(1973)
[165] McCloskey D.H., Hawkes S.J.. *J. Chromatogr.*, **13**, 1(1975)
[166] Lowry S.R., Ritter G.L., Woodruff H.S., Isenkour T.L.. *J. Chromatogr. Sci.*, **14**, 126(1976)
[167] Chastrette M.. *J. Chromatogr. Sci.*, **14**, 357(1976)
[168] Selzer R.B., Howery D.G.. *J. Chrometogr.*, **115**, 139(1975)
[169] Malinowski E.R.. *Anal. Chem.*, **49**, 612(1977)
[170] Weiner P.H., Liao L., Karger B.L.. *Anal. Chem.*, **46**, 2182(1974)
[171] Howery D.G., Weiner P.H., Blinder J.S.. *J. Chromatogr. Sci.*, **12**, 366(1974)
[172] Musumarra G., Scarlata G., Cirma G., Romano G., Palazzo S., Clementi S., Giuliette G.. *J. Chromatogr.*, **295**, 31–47(1984)
[173] DeLigny C.L., Spanjer M.C., Van Houwelingen J.C., Wessie H.M.. *J. Chromatogr.*, **301**, 311–324(1984)
[174] Fernandez-Sanchez E., Garcla-Dominuez J.A., Menendez V. ,Del Rio G.. *An. Quim., Ser A*, **82**, 507–512(1986)
[175] Weiner P.H., Parcher J.F.. *Anal. Chem.*, **45**, 302(1973)
[176] Weiner P.H., Howery D.G.. *Anal. Chem.*, **44**, 1189(1972)
[177] Warrew F.V. Jr., Bidlingmeyer B.A., Delaney M.F.. *Anal. Chem.*, **39**, 1890(1987)
[178] Warrew F.V. Jr., Bidlingmeyer B.A., Delaney M.F.. *Anal. Chem.*, **39**, 1897(1987)
[179] Howery D.G., Soroka J.M.. *J. Chemometrics*, **1**, 91–101(1987)
[180] Howery D.G., Soroka J.M.. *Anal. Chem.*, **58**, 3091–3095(1986)
[181] Howery D.G., Williams G.D., Ayala N.. *Anal. Chim. Acta*, **189**, 339–351(1986)

[182] Howery D.G., Soroka J.M.. *J. Chromatogr. Sci.*, **25**, 149–153(1987)
[183] Seaton G.G.R., Fell A.F.. *Chromatographia*, **24**, 208–16(1987)
[184] Kindsvater J.H., Weiner P.H., Klingen T.J.. *Anal. Chem.*, **46**, 982(1974)
[185] Weiner P.H., Howery D.G.. *Can. J. Chem.*, **50**, 448(1972)
[186] Zielinski W.L., Martive D.E.. *Anal. Chem.*, **48**, 1111(1976)
[187] Charton M.. *ChemTech.*, 245(1975)
[188] Hammett L.P.. *J. Am. Chem. Soc.*, **59**, 96(1937)
[189] Higman B.. *Applied Group-Theoretical and Matrix Methods*, Oxford University Press, Oxford: 1955
[190] Malinowski E.R., Ph.D. thesis, Stevens Institute of Technology, Hoboken, N.J.: 1961
[191] Weiner P.H.. *J. Am. Chem. Soc.*, **95**, 5845(1973)
[192] Kirkwood J.G., Westheimer F.H.. *J. Chem. Phys.*, **6**, 513(1938)
[193] Justice J.B., Isenhour T.L.. *Anal. Chem.*, **47**, 2286(1975)
[194] Rozett R.W., Petersen E.M.. *Anal. Chem.*, **47**, 1301(1975)
[195] Rozett R.W., Petersen E.M.. *Anal. Chem.*, **47**, 2377(1975)
[196] Rozett R.W., Petersen E.M.. *Anal. Chem.*, **48**, 817(1976)
[197] Rozett R.W., Petersèn E.M.. *Am. Lab.*, **9**(2), 107(1977)
[198] Burgard D.R., Perone S.P., Weibers J.L.. *Anal. Chem.*, **49**, 1444(1977)
[199] Wernimont G.. *Anal. Chem.*, **39**, 554(1967)
[200] Carey R.N., Wold S., Westgard J.O.. *Anal. Chem.*, **47**, 1824(1975)
[201] Edward J.T., Wong S.C.. *J. Am. Chem. Soc.*, **99**, 4229(1977)
[202] Deligny C.L., Van der veen N.G., Van Houwelingen J.C.. *Ind. Eng. Chem. Fundam.*, **15**, 336(1976)
[203] Fawcett W.R., Krygowski T.M.. *Can. J. Chem.*, **54**, 3283(1976)
[204] Howery D.G.. *Bull. Chem. Soc. Jap.*, **45**, 2643(1972)
[205] Duewer D.L., Freiser H.. *Anal. Chem.*, **49**, 1940(1977)
[206] Reeves R.L., Maggic H.S., Harkaway S.A., Meyers G.A.. *Inorg. Chem.*, **24**, 738–744(1985)
[207] Frans S.D., Harris J.M.. *Anal. Chem.*, **57**, 1718–21(1985)
[208] Ozeki T., Kihara H., Ikeda S.. *Anal. Chem.*, **60**, 2055(1988)
[209] Rao G.R., Zerbi G.. *Appl. Spectrosc.*, **38**, 795–803(1984)

[210] Shih L.B., Priest R.G.. *Appl. Spectrosc.*, **38**, 687-692(1984)
[211] Woodbury M.A., Cleland R.C., Hickey R.J.. *Behav. Sci.*, **8**, 347(1963)
[212] Swain C.G., Bryndza H.E., Swain M.S.. *J. Chem. Inf. Comput. Sci.*, **19**, 19(1979)
[213] Sneath P.H.A.. *J. Theor. Biol.*, **19**, 739(1976)
[214] Cammarata A., Menon G.K.. *J. Med. Chem.*, **19**, 739(1976)
[215] Menon G.K., Cammarata A.. *J. Pham. Sci.*, **66**, 304(1977)
[216] Weiner M.L., Weiner P.H.. *J. Med. Chem.*, **16**, 665(1973)
[217] Erickson B.C., Ruzicka J., Kowalski B.R.. *Anal. Chim. Acta*, **218**, 303-311(1989)
[218] Gauthier A., Zurli J., Cros B.C., Saries H.. *Rev. Eur. Etrud. Clin. Biol.*, **17**, 574(1972)
[219] John E.R.. *Osnovn. Probl. Elektrofiziol. Golovn. Mozga.*, **17**, 574(1972)
[220] Farber J., Tosovsky J., Hynck K.. *Act. Nert. Super.*, **16**, 258(1974)
[221] Bohidar N.R., Restaino F.A, Schwartz J.B.. *J. Pharm. Sci.*, **64**, 966(1975)
[222] Alpert D.J., Hopke P.k.. *Proc. Conf. Qulity Assurance Environ. Means.*, Denver, Colo.: Nov., 204, 1978
[223] Hopke P.K.. *J. Environ. Sci. Health*, **A11**(6), 367(1976)
[224] Hopke P.K., Gladney E.S., Gorden G.E., Zoller W.H., Jones A.G.. *Atoms. Environ.*, **10**, 1015(1976)
[225] Gaarenstroom P.D., Perone S.P., Moyers J.L.. *Environ. Sci. Technol.*, **11**, 795(1977)
[226] Knudson E.J., Duewer D.L., Christian G.D., Larson T.V.. in B.R. Kowalski, Ed., *Chemometrics: Theory and Applications*, ACS Symp. Ser. **52**, American Chemical Society, Washington, D.C.: 80, 1977
[227] Blifford I.H., Meaker G.O.. *Atmos. Environ.*, **1**, 147(1967)
[228] Peterson J.T.. *Atmos. Environ.*, **4**, 501(1970)
[229] Peterson J.T.. *Atmos. Environ.*, **6**, 433(1972)
[230] Lave L.B., Seskin E.P.. *Air Pollution and Human Health*, Johns Hopkins Universty Press, Baltimore, Md.: 33, 1977

[231] Linton P.W., Natusch D.F.S., Hopke P.K., Sdomon R.L.. *Proc. 4th Conf. Sensing Environ. Pollutants*, New Orleans, La.: Nov., 221, 1977
[232] Sanchez M.L., Casanova J.L., Ramos M.C., Sanchez J.L.. *Atmos. Environ.*, **20**, 53–56(1986)
[233] Koutrakis P., Spengler J.D.. *Atmos. Environ.*, **21**, 1511–1513(1987)
[234] Moranoll M.T., Daisey J.M., Lioy P.J.. *Atmos Environ.*, **21**, 1821–1831(1987)
[235] Lowenthal D.H., Rahn K.A.. *Atmos. Environ.*, **21**, 2005–2015(1987)
[236] Cobb G.P., Braman R.S., Gilbert R.A.. *Anal. Chem.*, **61**, 838–43(1989)
[237] Weiner P.H.. *ChemTech.*, 321(1977)
[238] Yoshizawa K., Ishikawa T., kinoshita M., Takeda A., Fujie L.. *Nippon Jozo kyokai Zasshi*, **69**, 581(1974)
[239] Reiner L., Piendle A.. *Brauwissenschaft*, **27**, 1(1974)
[240] Martens H., Sollbrg Y., Roer L., Vold E.. *Potato Res.*, **18**, 515(1975)
[241] Imbrie J.. Tech. Rep. No.6. ONR Task No. 389-135, Northestern University, Evanston: I11, 1963
[242] Dawson K.M., Sinclair A.J.. *Econ. Geol.*, **69**, 404(1974)
[243] Bolivar S.L., Campbell K., Wecksung G.W.. *J. Geochem. Explor.*, **19**, 723–743(1983)
[244] Roscoe B.A., Chen C.Y., Hopke P.K.. *Anal. Chim. Acta*, **160**, 121–134(1984)
[245] Magar M.E., Chuin P.W.. *Biophys. Chem.*, **1**, 18(1973)
[246] Hoesterey B.L., Windig W., Menzelaar H.L.C., Eyring E.M., Grant D.M., Pubmire R.J.. *ACS Symp. Ser.*, **376**, 189–202(1988)
[247] Sokal R.R., Snenth P.H.A.. *Principles of Numerical Taxonomy*, Freeman W.H., San Francisco: Chap.7, 1965
[248] Toerin D.F.. *Water Ress.*, **3**, 129(1969)