研究生教学用书

教育部研究生工作办公室推荐

化学计量学方法

Methods of Chemometrics

(第二版)

许 禄 邵学广 著

科学出版社
www.sciencep.com

研究生教学用书

教育部研究生工作办公室推荐

化学计量学方法

Methods of Chemometrics

（第二版）

许 禄 邵学广 著

科 学 出 版 社

北 京

内 容 简 介

本书为《化学计量学方法》的第二版，在第一版的基础上介绍了最新出现的研究方法，并收集了最新的研究成果。本书包括了化学计量学的主要内容，共分15章，分别为误差及数理统计基础、回归分析、相关分析和数据平滑、最优化方法、主成分分析和因子分析、偏最小二乘方法、多元校正及分辨、小波分析、遗传算法和模拟退火算法、人工神经网络法及在化学中的应用、模式识别方法、化合物结构表征和构效关系研究、组合化学、谱图库检索和结构解析专家系统及实验设计。此外，为方便读者查阅，本书还将常用数据信息列为附录。

本书可作为化学、生物化学、医学化学及环境化学等专业的研究生学习用书，也可供相关领域广大科技工作者参考。

图书在版编目(CIP)数据

化学计量学方法 / 许禄，邵学广著. —2版. —北京：科学出版社，2004

(教育部研究生工作办公室推荐研究生教学用书)

ISBN 7-03-013386-2

Ⅰ. 化… Ⅱ. ①许… ②邵… Ⅲ. 化学计量学-研究生-教材

Ⅳ. O6-04

中国版本图书馆CIP数据核字(2004)第044587号

责任编辑：刘俊来 王志欣 吴伶伶 / 责任校对：宋玲玲

责任印制：安春生 / 封面设计：陈 敬

科学出版社 出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

科源印刷厂 印刷

科学出版社发行 各地新华书店经销

*

1995年2月第 一 版 开本：720×1000 1/16

2004年9月第 二 版 印张：36 1/4

2004年9月第三次印刷 字数：692 000

印数：3 501—6 000

定价：54.00元

(如有印装质量问题，我社负责调换〈新欣〉)

前　言

化学计量学(chemometrics)是将数学和计算机科学应用于化学的一门新兴的交叉学科,是化学领域的一个重要分支。

数学是自然科学的语言,它在化学中的地位和作用日益突出和重要。特别是自20世纪70年代以来,随着计算机技术的迅速普及,数学和计算机科学在分析化学中应用日益广泛。于是化学计量学的内容得到了充实和扩展,使化学计量成为化学、生物化学、药物化学、环境化学和材料科学中信息处理的强有力手段。1974年,由美国的B. R. Kowalski和瑞典的S. Wold等发起,在美国华盛顿大学成立了国际化学计量学学会,开展了一系列的学术交流活动,促进了化学计量学的迅速发展。从1982年起,在美国分析化学杂志(*Anal Chem*)两年一度的基础评论中开辟了化学计量的专题;相继创刊了和化学计量学密切相关的一些杂志,如 *Chemometrics Intelligent Laboratory System*、*Journal of Chemometrics*、*Chemometric Window* 等。同时,国际上,一些化学计量学专著和系列丛书相继问世。

化学计量学是建立在多学科基础上的横向学科,在解决多种问题中显示了它的强大生命力。如在化合物定量结构-活性/性质相关性(QSAR/QSPR)研究中,在环境科学、过程分析和材料科学研究中等,化学计量已成为必不可少的重要组成部分。

近些年来,在化学计量学基础性研究的同时,有关化学计量学的应用工作也在大量开展。1994年,*Anal Chem* 所设"化学计量学"专题的评述中,检索的文献多达20 000篇,而1996年又增至25 000篇。

欧洲化学联合会的分析化学分部提出,在研究生的教学中,应该包括有四个方面的内容,即色谱学、光谱学、传感器和化学计量学。就是说,化学计量学在从它诞生至今的20多年中,作为分析化学中的一个分支已经得到公认。

在方法学方面,近年来,化学计量学也有很大发展,相继出现了许多新的方法,如人工神经网络法、遗传算法、小波分析及模拟退火等,这些方法得到了广泛的应用,并取得良好的结果。化学计量学已进入比较成熟的稳定发展阶段。可以期待,在化学、生物化学、医学化学、药物化学、环境化学和材料科学的发展进程中,将会发挥越来越重要的作用。

《化学计量学方法》一书出版于1995年,但是它文字材料的形成是在20世纪90年代初,已时过10年。鉴于化学计量学的发展,原书的一些内容需要更新,特

别是近年来发展起来的一些新的方法需要容纳进来。现在全书由原来的 11 章扩充为 15 章。第 1 章，尽管仍为误差和数理统计，但内容有了重要的修订和补充。第 2、3、5、6、7 和第 11 章为多元统计分析，内容包括回归分析、相关分析和数据平滑、主成分分析和因子分析、偏最小二乘、多元校正和分辨及模式识别。小波分析（第 8 章）是以小波变换为数学基础的信号分析技术，在 20 世纪 80 年代得到了迅速发展，在许多科学领域中得到了广泛应用。遗传算法和模拟退火算法（第 9 章）都是最优算法，但它们区别于单纯形和响应曲面法（第 4 章）。第 4 章和第 15 章主要用于实验的最优设计。结构解析的人工智能研究（第 14 章）是化学计量学很重要的组成部分，但是至今，在国内已有书籍中涉及的均较少。因此，我们把实验室中近 10 年来所取得的成果，分拓扑结构解析和立体结构解析两个部分进行了较详细的介绍。

另外，考虑到本书的读者对象主要是化学专业的读者，故在新版中仍然是不做数学上的详细推导和证明，只侧重介绍方法的概念和结论，同时尽可能给出应用实例，以使方法易读、易懂，在编写程序时也可作为参考。

中国科学院长春应用化学研究所的齐玉华、张庆友、王俊、董林和许即莲为本书的编写、定稿和程序的运行做了大量的工作，在此表示衷心感谢。

在本书出版之时，还要特别感谢科学出版社高等教育分社社长助理刘俊来先生、王志欣博士及吴伶伶编辑为本书出版所付出的辛勤劳动。

由于水平所限，书中缺点和错误在所难免，敬请读者批评指正。

作　者

2004 年于中国科学院长春应用化学研究所

目　录

第 1 章　误差及数理统计基础

因本书中有不少章要用到数理统计方面的知识，所以在这一章首先概括地介绍数理统计的基本概念，以利于其他章节的讨论。

1.1 误　　差

1.1.1 误差的定义

测量值 x 带有误差 E，测量值去掉误差就等于真值 μ_0，$\mu_0 = x - E$。所以误差的定义为

$$E = x - \mu_0$$

即测量值偏离真值的程度，也就是测量值的不确定度。

1.1.2 误差的类型

(1) 绝对误差

测量值大于真值时误差为正数，表示结果偏高；反之，误差为负数时表示结果偏低。这里的误差都是绝对误差，它具有与测量值和真值相对应的量纲。

(2) 相对误差

绝对误差在真值中所占的比率称相对误差，一般用百分率表示

$$\text{相对误差}(\%) = \frac{x - \mu_0}{\mu_0}$$

当真值为未知时，可用多次重复测定结果的算术平均值 $\bar{x}$ 代替 μ_0，相对误差没有量纲。

(3) 粗差

粗差也称过失误差，是由于非正常实验条件或非正常操作所造成的。如测量时对错了标志、误读了数码、实验仪器未达到预想的指标等。含有粗差的测量值常称为坏值或异常值，应予以剔除。

(4) 系统误差

系统误差是由某种原因所产生,并遵循一定的规律进行变化。例如,随样品或试剂用量的大小按比例进行变化。系统误差有一定的指向,例如称量一种吸湿性物质,其误差总是正值。从系统误差的来源看,它属于方法和技术问题,知道了产生的原因,便可消除或修正,所以此种误差也称可定误差。

(5) 随机误差

在相同条件下重复多次测定同一物理量时,误差大小或正负变化纯属偶然而毫无规律,这种误差称为随机误差,也叫偶然误差。随机误差单个地看是无规律性的,但就其总体来说,由于正负有相消的机会,随着变量个数的增加,误差的平均值将趋近于零。这种抵偿正是统计规律的表现,所以随机误差是可以用概率统计来处理的。

1.1.3 精密度和准确度

误差表示测量的不精密度和不准确度,即不确定度。精密度和准确度是两个不同的概念。精密度表示一组测定数据相互接近的程度或分散的程度,它的大小完全取决于偶然误差。在分析化学中,常用重复性(repeatability)和再现性(reproducibility)来表示精密度。重复性是指在完全相同条件下,即同一操作者、同一仪器、同一实验室,在较短时间内分析同一样品所得结果的精密度;再现性是指在不同的条件下,即不同的操作者、非同一台仪器、不同的实验室、不同的时间,但是用相同的分析方法和分析相同样品所得结果的精密度。准确度表示测量值与真值的偏离程度,它由系统误差和偶然误差共同决定。

如由 4 个学生用浓度准确为 0.1mol/L 的盐酸滴定浓度准确为 0.1mol/L 的氢氧化钠,氢氧化钠的体积准确为 10.00mL。每个学生重复测量 5 次,其结果示于表 1.1。

表 1.1 用盐酸进行氢氧化钠的滴定结果

学生	结果/mL	注释	学生	结果/mL	注释
A	10.08	精密但不准确	C	10.19	不准确也不精密
	10.11			9.79	
	10.09			9.69	
	10.10			10.05	
	10.12			9.78	
B	9.88	准确但不精密	D	10.04	准确而且精密
	10.14			9.98	
	10.02			10.02	
	9.80			9.97	
	10.21			10.04	

由表 1.1 可见，学生 A 尽管测试结果重复性较好，即精密，但是准确性较差（A 的均值为 10.10），所有结果均偏高，这是由于系统误差所致。学生 B 的测试落到准确值（即真值）的两侧，其均值为 10.01。此结果较准确，但精密度较差，主要受到了偶然误差的影响。学生 C 测量中既有偶然误差的影响，又有系统误差的影响，所以既不精密，也不准确。只有学生 D 测试结果比较精密（范围为 9.97～10.04mL），又比较准确（均值为 10.01）。

1.1.4　偶然误差的传递

(1) 线性加和

如 y 为测定量 a，b 和 c 等的线性组合

$$y = K + K_a a + K_b b + K_c c + \cdots$$

式中，K_a，K_b 和 K_c 等为常数。

则加和或差值的标准偏差是各量方差加和的平方根

$$\sigma_y = \sqrt{(K_a\sigma_a)^2 + (K_b\sigma_b)^2 + (K_c\sigma_c)^2 + \cdots}$$

如滴定中，移液管的初值和终值分别为 3.51mL 和 15.67mL，其标准偏差均为 0.02mL，则用去滴定液的体积及标准偏差分别为

$$\text{消耗的滴定液体积} = 15.67 - 3.51 = 12.16(\text{mL})$$

$$\text{标准偏差} = \sqrt{(0.02)^2 + (0.02)^2} = 0.028(\text{mL})$$

此例说明，组合的标准偏差大于单个读数的标准偏差，但小于各量的标准偏差之和。

(2) 乘除表达式

若计算 y 的表达式为

$$y = kab/cd$$

式中：a，b，c，d——测定量；

k——常数。

则相对标准偏差有如下关系

$$\frac{\sigma_y}{y} = \sqrt{\left(\frac{\sigma_a}{a}\right)^2 + \left(\frac{\sigma_b}{b}\right)^2 + \left(\frac{\sigma_c}{c}\right)^2 + \left(\frac{\sigma_d}{d}\right)^2}$$

如荧光的量子产率可用下式计算

$$\Phi = I_f/kcLI_0e$$

式中：I_0——入射光强度，$I_0 = 0.5\%$；

I_f——荧光强度，$I_f = 2\%$；

e——摩尔吸收，$e = 1\%$；

c——浓度，$c = 0.2\%$；

L——液槽长度，$L = 0.2\%$；

k——仪器常数。

则 Φ 的相对标准偏差为

$$r.s.d. = \sqrt{2^2 + 0.2^2 + 0.2^2 + 0.5^2 + 1^2} = 2.3(\%)$$

由此可见，最终结果的相对标准偏差略大于上述分量中具有最大相对标准偏差的那个分量(I_f)。这一结果给我们的启示是，若拟提高测试的精度，则首先应该设法改善具有最大相对标准偏差的那个分量的测试精度。

另外，对于某一量的乘方，如

$$y = b^n$$

则 y 的相对标准偏差为

$$\frac{\sigma_y}{y} = \left| \frac{n\sigma_b}{b} \right|$$

因为 b 和 b^n 不是分别独立的量。

(3) 其他函数

若 y 是 x 的函数

$$y = f(x)$$

则 x 和 y 的标准偏差具有如下关系：

$$\sigma_y = \left| \sigma_x \frac{\mathrm{d}y}{\mathrm{d}x} \right|$$

如某溶液的吸收值 A 为光透过率的函数

$$A = -\lg T$$

若 T 的测定值为 0.501，标准偏差为 0.001，则 A 的值及其 $\mathrm{d}A/\mathrm{d}T$ 分别为

$$A = -\lg 0.501 = 0.300$$

和

$$\frac{\mathrm{d}A}{\mathrm{d}T} = -\lg e/T = -0.434/T$$

由此可得 A 的标准偏差为

$$s = |0.001 \times (-0.434/0.501)| = 0.000\ 87$$