科学版研究生教学丛书

化学计量学

一些重要方法的原理及应用

许 禄 主编

科学出版社

北京

内 容 简 介

本书介绍了化学计量学中的一些重要方法。全书共 10 章。其中,第 1 章概略地回顾了化学计量学的发展历史及现状;第 2～4 章介绍了主成分分析及在此基础上的偏最小二乘方法、多元分辨方法和三线性分解;第 5 章介绍了人工神经网络法(目前在化合物构效关系研究中应用广泛);第 6 章介绍了遗传算法及模拟退火算法(它们是近年发展起来的寻优算法);第 7 章介绍了小波分析(主要用于信息压缩、重叠峰分解及基线校准等);第 8 章介绍了模式识别中最新方法的发展;第 9 章及第 10 章均为化学计量学方法在光谱中的应用。

本书可供生物化学、药物化学、毒物化学、医学化学及环境化学等专业的研究生和相关工作人员阅读参考。

图书在版编目(CIP)数据

化学计量学:一些重要方法的原理及应用/许禄主编.—北京:科学出版社,2004

(科学版研究生教学丛书)

ISBN 7-03-011643-7

Ⅰ.化… Ⅱ.许… Ⅲ.化学计量学-研究生-教学参考资料 Ⅳ.O6-04

中国版本图书馆 CIP 数据核字(2003)第 055478 号

策划编辑:刘俊来 王志欣 / 文案编辑:吴伶伶

责任校对:包志虹 / 责任印制:安春生 / 封面设计:曹 烨

科学出版社发行 各地新华书店经销

*

2004 年 2 月第 一 版 开本:B5(720×1000)

2004 年 2 月第一次印刷 印张:19

印数:1—3 000 字数:359 000

定价:20.00 元

(如有印装质量问题,我社负责调换〈路通〉)

《化学计量学——一些重要方法的原理及应用》
编　委　会

序

2001 年在长沙由许禄教授主持的有关国家自然科学基金化学计量学重点项目评议会上,与会专家提出,化学计量学基础研究在我国取得了可喜的成果,但许多研究成果还停留在论文或实验阶段,未见付诸应用。另外,在方法的掌握上还有欠缺。如何进一步提高我国化学计量学的水平,由此来推进化学计量学在生产科研中实际应用,是发展这一分支学科十分迫切的任务。在国家自然科学基金委员会的支持与鼓励下,与会专家商定,2002 年在北京举办化学计量学讲习班。这期在首都师范大学举行的讲习班取得了很好的成效。科学出版社拟将讲习班各讲座内容汇集成书出版。许禄教授希望我为本书作序,其实,在我与梁逸曾教授合写的第 1 章中,想说的话都写进去了,再写序言已无多大必要。期望通过该书的出版,将讲习班的效果进一步扩大,更多的讲习班将会举办。最重要的是化学计量学能为更多的化学工作者及其他行业的同仁所认可,并在他们的工作中得到实际应用。我想,这也是主编者和作者们,以及所有从事化学计量学研究工作的同仁们所热切期望的。

俞汝勤

2003 年 7 月于湖南大学

前　言

化学计量学是将数学、计算机科学应用于化学的一门新兴的交叉学科，是当代化学领域的一个重要分支。

化学计量学是于 1974 年由美国学者 Kowalski 和瑞典学者 Wold 共同发起的，他们在美国华盛顿大学成立了国际化学计量学学会。迄今，已近 30 年时间。几十年来，化学计量学不仅在方法学方面有很大的发展，而且化学家们做了大量的、广泛的应用，对于化学学科的发展起到了很大作用。特别是近些年来，在化学计量学领域发展起许多新的方法，如人工神经网络法、遗传算法、模拟退火算法及小波分析等。大凡有用的算法一经提出，立即就会引起人们的广泛关注。如人工神经网络法，在一年之内，有关的文章可以多至万篇。再如，遗传算法和小波分析，直到现在，每年涌现的文献量都是非常多的。

然而，从大量的杂志或专业会议的审稿中我们发现，特别是国内，在一些方法的应用上存在很多问题。其问题来源主要是对方法的理解有欠缺，因此在实际的应用上比较混乱。例如，人工神经网络法是建立在现代神经科学研究成果基础之上的一种抽象的数学模型，它反映了大脑功能的若干基本特征，如自学的功能。同时，人工神经网络法既可以作为模式识别器，又可构造数学模型以用于精确值的计算。由此，在化学、生物化学、药物学、药物化学、医学化学等领域得到极为广泛的应用。但某些模型，在理论上，如反向传输人工神经网，初始权重的设置问题迄今未能解决；在使用上，如过拟合和过训练，是很易发生的。由于这些问题的存在，会导致构造的数学模型不稳定，如在定量结构-活性/性质相关性(QSAR/QSPR)的研究中，使预测的效果较差。在其他的一些方法如遗传算法中，问题的表现形式不一样，但其要害均是由于对方法的理解深度不够。

为了提高我国化学计量学的水平，以使化学计量学更好地在科研和生产中发挥实际效用，特邀请了我国著名的化学计量学专家，如俞汝勤院士等，分为专题，从方法的原理到方法的应用，一一进行较为详细的讲解。所讲专题都是目前化学计量学中的重要方法，或者在理解上比较困难的方法。这些专题，在 2002 年北京的化学计量学学习班上进行了宣读，获得普遍好评。现特编辑出版，以飨读者。

本书的特点是侧重一些重要方法的介绍，而不是面面俱到地、系统地进行讲解，这样，特别是对已经有了一定基础的人是非常有裨益的。全书共 10 个专题，编为 10 章。其中，化学计量学：发展与展望(第 1 章，俞汝勤、梁逸曾)，概略地回顾了化学计量学的发展历史及现状。主成分分析是许多方法的基础，本书介绍了偏最

小二乘方法(第 2 章,吴海龙)、多元分辨方法(第 3 章,梁逸曾)及三线性分解(第 4 章,吴海龙)等。第 5 章(许禄)是人工神经网络法,前边已有介绍,不再赘述。遗传算法及模拟退火算法(第 6 章,李通化、蔡文生)是近年发展起来的寻优算法,特别是遗传算法用得非常广泛。小波分析(第 7 章,邵学广)是一个新的数学分支,有人说它是近年来工具及方法上的重大突破,是泛函分析、傅里叶分析、样条分析、调和分析和数值分析的完美结晶。第 8 章(梁逸曾)介绍了模式识别中最新方法的发展。第 9 章(吴玉田)及第 10 章(张卓勇)均为化学计量学方法在光谱中的应用。其中,褶合光谱法的理论与实践(第 9 章)侧重的是近红外光谱;第 10 章为化学计量学方法在光谱干扰校正中的应用。曹同成博士参加了遗传算法的编写。

在本书出版之际,我们要特别感谢国家自然科学基金委员会在资金上给予的资助。同时要特别感谢首都师范大学为学习班成功地举办所提供的方便、所付出的辛劳和努力。作者还想衷心感谢科学出版社高教分社社长助理刘俊来先生、王志欣博士及吴伶伶编辑所付出的辛勤劳动。

由于编者水平有限,书中缺点和错误在所难免,敬请读者不吝赐教。

许 禄

于中国科学院长春应用化学研究所

2003 年 6 月

目　　录

第 1 章　化学计量学:发展与展望

1.1　Wold 的命名与学术界的反响

微型计算机于 20 世纪 70 年代初开始大量进入化学实验室,孕育着一门反映化学学科信息化的化学新分支学科的诞生。从瑞典 S. Wold 为他的基金项目构思出“化学计量学”(chemometrics)这个独特的名称至今,以这个名称命名的化学分支已走过了 30 多年的历史。S.Wold 研究采用统计学、应用数学和计算机科学的方法来优化化学实验、化工生产和化学量测过程,并从化学量测的数据中最大限度地提取相关信息。美国 B. Kowalski 赞同 S. Wold 提出的概念,他们于 1974 年共同发起成立了国际化学计量学学会,至今已发展成为一个在许多国家拥有分会的学术团体,该学会及其分会与相关学术机构出版了不少带有化学计量学一词的学术刊物,如 *Journal of Chemometrics*、*Chemometrics and Intelligent Laboratory Systems*、*Chemometric Window* 等。在一些分析化学期刊如 *Analyst*、*Analytica Chimica Acta*、*Chromatography A*、*Trends in Analytical Chemistry* 等设有 chemometrics 专栏。美国化学学会主办的 *Analytical Chemistry* 从 80 年代起推出 chemometrics 的双年综合评论。一些化学化工期刊如 *Chemical Information and Computer Science*、*Chemical Engineering and Computer Science* 等均发表了大量有关化学计量学理论与应用的论文。显然,国际学术界已普遍接受了S.Wold 提出的命名。

1.2　化学计量学与分析化学

化学计量学的发展与分析化学有密切的关系。在创立这一化学分支的过程中,分析化学家做出了重大的贡献。分析化学是化学量测与表征的科学,Valcarcel 将其定义为“发展、优化、应用量测过程,以获取全局或局部性化学品质信息,解决有关量测课题的计量科学”。S.Wold 强调:他提出“化学计量学”的初衷即是从化学量测数据中获取、表述、显示相关化学信息。当年,他就是从“化学数据分析”、“化学中的计算机”、“化学计量学”等备选名称中确定他的基金申请项目名称的。从分析化学的角度我们认为,化学计量学的研究对象是化学量测的基础理论与方法学。它构成分析化学第二层次基础理论的重要组成部分。这里我们将化学学科各门二级学科共有的基础理论如理论化学等归为第一层次的基础理论,分析化学

本身区别于其他化学二级学科的基础理论列为第二层次，而分析化学所属各分支的基础理论如穆斯堡尔谱学理论、圆二色谱理论等列为第三层次。前两层次是任何分析化学家均应掌握的，第三层次则因方向不同而异。化学计量学为分析化学在严谨的数理基础之上构建采样、检测、校正、分辨、识别、误差、优化等基础理论做出了重要贡献。那种认为分析化学没有基础理论，甚至因而不能将其认作为一门独立学科的论点显然是一种偏见。

1.3 FECS 的 Eurocurriculum

化学计量学的发展已对分析化学产生了深刻的影响。特别值得注意的是，化学计量学能为分析仪器的智能化提供新理论和新方法，为新型高维联用仪器的构建提供新思路和新方法，这是 21 世纪分析仪器向软件化发展的新突破口。此外，随着微型计算机的飞速发展，化学波谱数据库的建立与检索及化学人工智能和专家系统的研究也将取得长足进步。在采用计算机网络技术将多种波谱仪器连接的基础上，将数值化计算技术（这是近年来化学计量学研究所采用的方法）与基于经验的逻辑推理方法的有机结合，可望解决化合物结构自动解析的难题，并使长期困扰分析化学家的混合物波谱同时定性定量解析成为可能。在分析化学领域中，化学计量学的发展前景是广阔的。

欧洲化学会联合会（FECS）的分析化学分部（WPAC）在综合 180 所欧洲大学的分析化学课程内容基础上，推出了《欧洲分析化学本科教学大纲》（*Eurocurriculum of Analytical Chemistry*），并以此为基础出版了《分析化学教程》。这个大纲包含总论、化学分析、物理分析和计算机分析化学四大部分，第四部分的主体内容是化学计量学。在研究生教育方面，WPAC 认为，研究生的教学内容可因学校与导师的方向不同而异，但应涵盖分析化学的四大支柱，即色谱学、光谱学、传感器和化学计量学。FECS 除包含欧洲大约 30 个国家的化学会组织外，还有美国、澳大利亚、埃及、日本与中国的观察员参与。可以说，至少化学计量学在分析化学领域中的地位已得到了较正式的公认。

1.4 属于整个化学的学科分支

化学计量学与分析化学关系密切，但它的研究对象并不局限于分析化学。我们将化学计量学认作一个化学分支领域，它属于整个化学学科。例如，化学计量学中的 QSAR（定量构效关系）研究物质的化学结构与其活性或其他性质之间的定量关系，而这实际上是化学研究最基础的问题。无机化学中有关配位化合物的结构与性能关系研究，导致了许多无机化学核心理论的构建。有机化学中有关有机化

合物自由能线性相关的研究,被认作是 QSAR 研究的前身。

化合物的结构与性能关系的研究及相关立体化学等理论的建立,推动了建立在这些理论之上的药物化学等相关学科的发展。以有机合成或自动化、智能化为研究对象的合成计量学(synthometrics)这一独特的化学计量学分支,可能对有机化学的发展产生重要的影响。

物理化学作为化学的理论分支,其核心的研究对象也是化合物结构与物理化学性质之间的关系。目前,化学计量学的应用已涵盖许多应用化学分支,如化学工程、环境化学、食品化学、农业化学、药物化学等。

1.5　“算法驱动”与“问题驱动”

关于什么是一门学科的基础理论,往往存在一些误解。化学计量学是一门交叉学科,化学计量学为化学量测提供理论和方法。化学计量学研究着眼于发展化学数据解析的新理论和新方法及其在各个化学分支的应用研究。

化学计量学家建立了一系列独特的多元校正、多元分辨及模式识别方法,如 SMICA、秩消失因子分析、渐进因子分析、直观推导演进特征投影法等。统计学、应用数学与计算机科学的发展又为化学计量学提供了诸如基于自然计算的进化算法、小波分析等新工具。

从化学的角度看,化学计量学研究从根本上讲是“问题驱动”的。脱离化学实际,单纯进行有关纯粹统计学、数学等方面的研究受到 S.Wold 的批评。S.Wold 从其父(统计学家 H.Wold)那里学习了偏最小二乘算法并成功地引入化学研究。他本人最有条件,但也是最反对化学家进行脱离化学实际的基础数学研究的。统计学、数学问题本身的基础研究应是统计学家与数学家之长,对化学家而言,开展这种研究可能是舍其所长而取其短,难免导致一些低水平重复工作,这是应当避免的。基础科学需要实际问题的驱动力推动其发展,前面提到的图论的发展与有机化学的关系就是一个例子。这种研究当然是具有创新意义的,不能认为是一般的“知识整合”。难道 DNA 双螺旋结构模型的提出只是一个结构化学的例行研究案例?显然我们主要从它对生物医学发展的划时代意义来进行学术评价。

化学计量学在化学各个分支科学中的应用研究取得了重要成果,可以说是“问题驱动”研究的实例。环境化学中的污染源识别,被称为环境计量学(envirometrics);商品防伪辨识及各种商品的质量检测,被称为品质计量学(qualimetrics);药物化学中分子设计、新药发现及结构性能关系(QSAR),被称为药物计量学(pharmacometrics)。化工过程分析、工艺过程诊断、控制和优化是化学计量学在工程科学中应用的实例,形成了诸如多参量过程控制(multi-parameter process monitoring)等独特方向。

如前一节所述,化学计量学主要是应用数学、统计学与计算机科学的工具和手段及其最新成果来设计化学实验、优化化工生产和量测过程,并通过解析化学量测数据以最大限度地获取化学及其相关信息。这样,自然就产生了一个这样的问题,作为一个学科,怎样来体现它的创造性、原始性问题?

有人认为,化学计量学大量应用数学家、统计学家及计算机科学家的成果来解决化学中的问题,至多只属于知识整合性工作,难于做出具有创造性和原始性的工作来,故对化学计量学抱有悲观看法。

化学计量学家在开展"算法驱动"的基础研究时,可能被认为这是化学家在做统计学、数学或计算机科学家的事。持这种观点的人不承认涉及解决化学问题的这种研究工作属于主流化学范畴。其实,化学家公认的理论化学基础研究运用薛定谔方程等数学方法和量子力学方法,能不能说这些工作是"数学或力学家的事",不属主流化学范畴呢?

现代科学学科交流、协同发展的趋势,实际上已完全不能纯粹从经典学科划分的角度来看待这类问题。在理论化学与化学计量学中得到重要应用的图论,其本身的发展就包含了有机化学等化学家的重要贡献。把图论的基础研究与涉及化学应用的研究严格划分实际上是很困难的。Wilkins 通过对脱氧核糖核酸(DNA)分子的 X 射线衍射研究,证实了 Watson 与 Crick 提出的 DNA 反向平行双螺旋模型。这些高水平的研究工作看起来似乎纯粹是化学家的事,上述三位学者的成果为合理解释遗传物质的功能与生物的遗传与变异奠定了理论基础,成为生物与医学领域里程碑式的发现,因而获得诺贝尔生理医学奖。这是化学与生物医学学科交缘发展的光辉事例。

从学科发展来看,化学计量学研究中不应在进行"算法驱动"型研究时忽视"问题驱动"的研究。

1.6 化学数据挖掘的大好机遇

化学是一门实验的科学,其最基本的研究对象是物质的组成、结构和性质及其相互联系和变化的规律。目前已知化合物的数量已超过两千多万种,其中包含的化学知识(信息)量可能只有生物学科积累的基因序列能与之媲美(基因序列数据也是化学数据),而这些化学信息基本上是通过实验获得的。在长期的化学实践中,积累了海量的化学信息,这些信息散布在浩如烟海的各类化学出版物中。虽然这些化学信息为人们探索自然界的奥秘提供了基础,但因为数据量的迅猛增加却造成了使用上的困难,常规手段已无法满足化学家的需要,因此众多的化学数据库应运而生。近年来,人们在利用数据库对化学数据进行研究时,逐步认识到海量数据的利用十分困难,且不充分,更具价值的规律性的信息和知识还隐蔽在数据内

部。如何从化学数据中发现更多、更有价值的化学规律正逐步成为化学家们关注的焦点,徐光宪先生曾指出:“从科学发展史看,科学数据的大量积累,往往导致重大科学规律的发现。”从 19 世纪 60 年代化学积累的数 10 种元素和上万种化合物的数据中,门捷列夫于 1869 年发现了元素周期律。20 世纪 30 年代,化合物累积到 100 多万种,后来鲍林发表《论化学键本质》。到 20 世纪末,*Chemical Abstract* 登记的分子、化合物和物种的数目已超过 2340 万种,比鲍林总结化学键理论时扩大了 10 余倍,而化学家似乎还没有充分利用这一化学宝库来总结规律。化学计量学的数据挖掘方法面对的正是这一难得的机遇。可以预期,化学计量学方法的有效运用,将促进化学家发现新规律的工作向前迅速推进。

1.7　化学计量学与生物信息学

由于 DNA 结构的揭示与人类基因组研究等方面的进展,化学与生命科学的交流更趋密切。在涉及化学所面临的大量数据积累的同时,我们很自然地联想到在化学的协助下,生命科学向前发展所面临的形势。这可以从与化学计量学有很多相似之处的生物信息学(bioinformatics)的发展来进行考查。生物信息学这一生物学分支,是从蛋白质结构、分子进化、DNA 排序等研究工作逐步发展起来的,计算机在这些研究工作中起到了重大的作用。生物学家与化学家协同数学、物理和计算机科学家逐步形成了生物信息学的学术群体。1985 年出版的 *Computer Applications in Biosciences* 刊登大量生物信息学论文,后来这一期刊改名为 *Bioinformatics*。这一生物学学科分支在利用因特网在全世界交流序列信息方面十分成功。从基因测序来考查生物物种逐渐成为生物学家的例行思维模式。从广义角度而言,生物信息学被认为是信息技术在生物数据的管理与分析中的应用,它涵盖诸如人工智能、机器人及基因分析等内容;就基因分析而言,生物信息学这一名词曾被用于概括生物(DNA/蛋白质)序列数据的处理与分析这样的研究范畴。显然,生物信息学的某些侧面与 S.Wold 当年将“化学数据分析”或“化学中的计算机”命名为“化学计量学”颇为相似。生命科学面临的数据爆炸与化学家所面临的形势颇为相近且大大超越,而且应当记取,基因序列与蛋白质结构等数据与信息就其本质而言也是化学数据与信息!与化学学科相似,生命科学可能也是处在出现新的发现甚至新的突破的前夜。化学家十分关切生命科学的进展,化学计量学也应关注生物信息学的发展。有可能出现这两个学科分支的协同研究课题,两个学科分支能为化学与生物学之间搭建桥梁,化学计量学更接近化学,生物信息学更接近生物学,因而各有所长,而两者的方法体系则有不少相似之处。有人建议将化学计量学更名为化学信息学(chemoinformatics)。生物计量学有较特定的发展背景。Mount 出版的一本巨著 *Bioinformatics* 用了 *Sequence and Genome Analysis* 的副

标题,这意味着基因序列在生物信息学中的主体地位。本书所讨论的化学计量学在化学领域的研究范畴,似乎较目前公认的生物信息学在生物学领域的研究范畴有较大差异。更改特别是取消名称或仿效生物信息学来构思对应的目前化学计量学的研究范畴,可能导致某些混乱。如果另从最广义的"信息"概念来定义"化学信息学",将一般文献检索或化学文献学、因特网技术(这在生物信息学中有特定的重要意义,因特网是传输某些特大容量基因序列数据的有效手段)等均列入化学信息学范畴,则本书论及的化学计量学的一些研究重点反而可能被冲淡,以致失去此学科分支的特色。其实在生物领域,也存在"生物计量学"(biometrics)分支,他们仍区别于生物信息学而存在。化学计量学在方法学体系的构建方面,进展似乎超过了生物计量学。不排除在化学学科中让化学信息学与化学计量学并存,但有无必要取消化学计量学而一概称为化学信息学(涉及生物计量学与生物信息学,在生物界并无类似之举),似乎宜仔细斟酌。

1.8 展 望

化学计量学正在展示十分诱人的发展前景。环境计量学将为可持续发展做出重要的贡献。品质计量学是提升国家经济水平的重要手段。药物计量学在新药设计、中草药国际化方面,展示了良好的发展前景。定量构效关系在新材料设计中的应用已引起材料科学家的重视。多参量过程控制可能成为化工及其他生产过程控制的重要方向。合成计量学提出了要从根本上改变有机合成单纯劳动密集型手工操作约束其大发展的局面。化学计量学与量子化学计算相结合,将使一些暂时无法进行理论运算的复杂体系有可能应用量子化学理论成果进行研究。化学计量学作为分析仪器智能化的基础,可望在新一代分析仪器构建中发挥巨大作用,为振兴民族分析仪器产业做出贡献。

经历了 30 余年的发展,化学计量学在 21 世纪进入了新的发展年龄段。莎士比亚在其名剧"皆大欢喜"中论及人的从婴儿、学童到情人、战士……的各个年龄段。化学计量学已不再是"婴儿"或"学童"。在"化学计量学热"中认为这一方向能解决一切化学难题的"情人"视角,也已趋冷静,化学计量学进入了稳定深入发展的较成熟的阶段。期望化学计量学在方法学研究方面取得新的进展,特别是在复杂的化学与分析化学问题包括生命与材料科学中的化学问题的解决方面,取得更多新的成果。

第 2 章　主成分分析

2.1　引　　论

主成分分析在文献[1～14]中有许多同义词，如特征值分解(eigenvector decomposition)、奇异值分解(singular value decomposition)、抽象因子分析(abstract factor analysis)、主因子分析(principal factor analysis)、特征矢量投影(eigen vector projection)、K-L 投影(K-L projection)等。其基本含义为利用特征分析的数学方法对数据阵求取特征值和特征矢量。方法是将原变量进行转换，使数目较少的新变量成为原变量的线性组合，而且新变量应最大限度地表征原变量的数据结构特征，同时不丢失信息。也即主成分分析的目的就是将数据降维，以消除众多信息共存中相互重叠的信息部分。对一个矩阵，利用其变量之间的共线性，对数据进行简约。这样，可直观显示(图示)其内在特性，可提取抽象因子有助于对相互关系的简明解释，可有效克服不稳定算法因严重共线性即病态所引起的计算误差的放大。

2.2　主成分分析的理论

主成分分析涉及多维空间中的投影概念。为了便于理解其意义，以二维空间中的主成分分析为例说明。

2.2.1　二维空间的主成分分析

利用气相色谱(GC)得到的 8 个样品中苯和二甲苯的含量如表 2-1 所示。

表 2-1　8 个样品中苯和二甲苯的含量

序　号	B	T	Bmc	Tmc
1	48	26	13	12
2	44	20	9	6
3	40	24	5	10
4	38	18	3	4
5	32	9	−3	−5
6	28	6	−7	−8
7	26	5	−9	−9
8	24	4	−11	−10
平　均	35	14	0	0

注：B 表示苯；T 表示二甲苯；Bmc 和 Tmc 为减去平均值后的值。

它可看成是在二维空间中的一组测试点(y_{1i}, y_{2i})(i=1,2,…, m),这里,m 为样品数,m=8。如果将二维数据降至一维数据,实际上,就是将二维空间的点投影到一维空间的一条线上。如果没有任何约束条件,其投影的方向将有无穷多个。但主成分分析操作采用的投影方向的约束条件是,在一维空间中的这条直线必须包含了原数据的最大方差,即沿着这条直线,使原数据的方差达到最大。如图 2-1 中点 i(i=1,2,…,8),向直线 P_1 投影为点 i'(i'=1,2,…,8),这些点的重心为 O,其分布可用它们以中心点 O 的距离的平方和表示。原数据点的距离分布为

$$S^2 = |O1|^2 + |O2|^2 + \cdots + |O8|^2$$

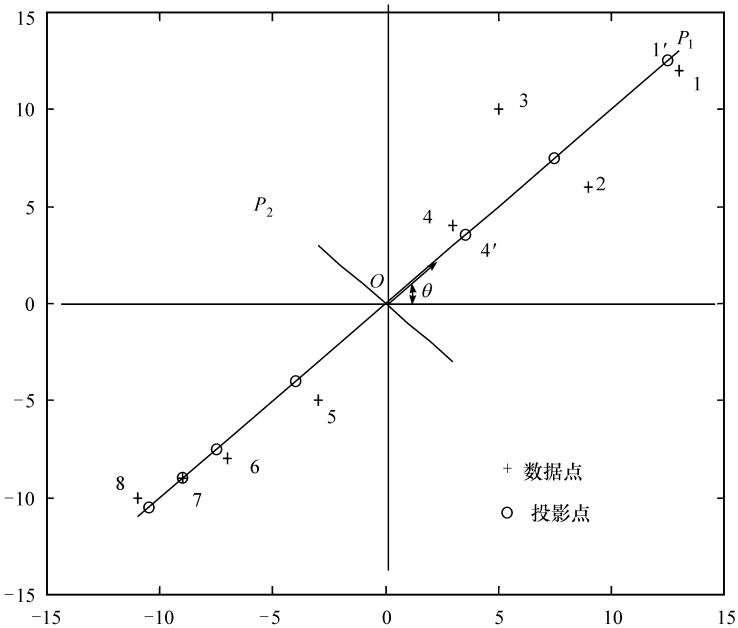

图 2-1 Bmc 和 Tmc 的二维图示及特征矢量

如果用在 P_1 上的投影点表示,则 $|Oi|^2 = |Oi'|^2 + |ii'|^2$,所以有

$$S^2 = |O1'|^2 + |O2'|^2 + \cdots + |O8'|^2 + |11'|^2 + |22'|^2 + \cdots + |88'|^2$$

主成分分析选择投影直线 P_1 使上式中 S^2 的值最大。这条直线也正好是这些原数据点的最佳拟合线,它使所有的原始数据点到在 P_1 直线上对应投影点垂直距离的平方和最小。通常,P_1 称为主成分空间,图 2-1 中箭头表示该空间中的单位向量,即载荷向量。如点 1 和点 7 在 P_1 空间中的投影点分别为 1′和 7′,它们在 P_1 空间中的坐标分别为 t_1 和 t_7,即 P_1 空间中用载荷向量对投影点距重心点距离度量的得分。

上述例子中,使用一维新变量 P_1 表征二维的原数据(y_{1i}, y_{2i})(i=1,2, …, m)

的结构特征，新变量包含了原数据中的绝大部分信息特征，可称它为第一主成分。还有部分剩余的信息没有被包含进来，可以使用与选取第一主成分相同的方法，再选出第二主成分来描述这剩余信息部分。第二主成分应在与第一主成分不相关的其余变量中能包含最大的方差。对于多维空间，依次类推，可以选出第三、第四等主成分。

在主成分分析中，对原变量进行了转换，即新变量是原变量的线性组合，原坐标系的原点经一简单转换后，放到重心 O 处（主成分分析关心的是数据系列所包含的信息，而且它使数据运算误差更小）。根据几何规则，新变量很容易被原数据以线性组合形式表示出来（如下式所示）。值得指出的是，P_1 及 P_2 的计算式中的系数 a 和 b 有约束条件 $a^2+b^2=1$ 和 $c^2+d^2=1$。使用矩阵形式表示

$$\begin{bmatrix} p_1 \\ p_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$p_1 = ax_1 + bx_2 = x_1\cos\theta + x_2\sin\theta$$

$$p_2 = cx_1 + dx_2 = x_1(-\sin\theta) + x_2\cos\theta$$

$$x_{1i} = y_{1i} - \bar{y}_1 \qquad x_{2i} = y_{2i} - \bar{y}_2$$

特征矢量（主成分 1）

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$$

特征矢量（主成分 2）

$$\begin{bmatrix} c \\ d \end{bmatrix} = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$$

2.2.2　多维空间中的主成分分析

对于 m 维空间中的主成分分析，将新变量 $p_1, p_2, \cdots, p_n$ 表示为原变量 $x_1, x_2, \cdots, x_m$ 的线性组合

$$\begin{aligned} p_1 &= v_{11}x_1 + v_{12}x_2 + \cdots + v_{1m}x_m \\ p_2 &= v_{21}x_1 + v_{22}x_2 + \cdots + v_{2m}x_m \\ &\cdots \quad \cdots \quad \cdots \quad \cdots \\ p_n &= v_{n1}x_1 + v_{n2}x_2 + \cdots + v_{nm}x_m \end{aligned}$$

则

$$\boldsymbol{V} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & & \vdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix}$$

$$\boldsymbol{P}=\begin{bmatrix}p_1\\p_2\\\vdots\\p_n\end{bmatrix}\qquad \boldsymbol{X}=\begin{bmatrix}x_1\\x_2\\\vdots\\x_m\end{bmatrix}$$

$$\boldsymbol{P}=\boldsymbol{V}\boldsymbol{X}$$

约束条件如下。

1) 对于任意两个主成分,有

$$v_{j1}\,v_{i1}+v_{j2}\,v_{i2}+\cdots+v_{jm}\,v_{im}=0$$

2) 对任一主成分,有

$$v_{i1}^2+v_{i2}^2+\cdots+v_{im}^2=1$$

表明在 m 维空间内,新变量的系数是一 m 维矢量。各矢量互为正交的,矢量为单位长度。

可以将新变量的协方差阵表示为原变量的协方差阵

$$C_p=\sum_{j=1}^{m}\sum_{k=1}^{m}v_{1j}\left(\frac{1}{n}\sum_{i=1}^{m}x_{ij}\,x_{ik}\right)v_{1k}$$

$$=\sum_{j=1}^{m}\sum_{k=1}^{m}v_{1j}\;c_{x,jk}\,v_{1k}=VC_xV'$$

约束条件(2)表明 $v'v=1$。

根据拉格朗日乘法,左乘 v_1' 有

$$C_x\,v_1=\lambda_1\,v_1$$

$$v'_1\,C_x\,v_1=v'\lambda_1\,v_1=\lambda_1(v'_1\,v_1)=\lambda_1$$

式中,λ_1 为特征值,对应的矢量为特征矢量 P_1,称 P_1 为主成分。

根据约束条件,用 $v_2,\cdots,v_m$ 计算,可以得到

$$\lambda_2,p_2,\cdots,\lambda_m,p_m$$

在 m 空间内,可得 m 个主成分,相对应 m 个特征值

$$\lambda_1,\lambda_2,\cdots,\lambda_m$$

第一主成分包含了最大偏差量,第二主成分次之,依次类推。一般取前几个主成分,即将多维空间数据降低为低维空间数据。取主成分个数 l 参考下式

$$T(\%)=\sum_{i=1}^{l}\lambda_i\Big/\sum_{i=1}^{m}\lambda i$$

一般 T 高于 80%。在定量分析中,可根据问题的要求,取靠近 100%的 l 值。值得注意的是,不同变量间的数据差异较大或不同变量的单位不同时,计算前,应进行标准化处理,即变量与均值之差被标准差除。也就是说,使用相关矩阵替代上述的协方差矩阵。对于吸光度测量数据,因单位相同,也可不做数据的标准化处理,这样显得简洁明快。

2.3　主成分分析算法[5~14]

2.3.1　特征值分解

总结上述计算过程是，先进行数据的预处理，得矩阵 X，再计算协方差矩阵 Z，然后根据协方差矩阵计算特征值和特征矢量。这一方法称为特征值分解方法。这一计算步骤在 MATLAB 语言环境中，仅需一个语句[V, D]=eig(X)就可得到特征值对角阵 D 和满秩正交矢量阵 V，且 $XV=VD$。

2.3.2　奇异值分解

奇异值分解也是一种对数据矩阵直接进行分解的方法。它性能稳定获广泛好评。利用 MATLAB 语言对数据矩阵 Xnm 作奇异值分解，[U, S, V]=svd(X)，可得

$$X = USV'$$

式中，U、S 和 V 的大小分别为 $n\times r$、$r\times r$ 和 $m\times r$，且 $U'U=I_{r\times r}$，$V'V=I_{r\times r}$，$X'X=VS^2V'$，$XX'=US^2U'$。

与下节中的 $X=TP$ 相比，可知 $T=US$，$P=V'$ 且 $\lambda=S^2$，即实数矩阵的特征值等于相应奇异值的二次方。由于 svd 性能优异且表示简洁，已被广泛采用。

2.3.3　NIPALS 法[10]

计算主成分的方法还有非线性迭代偏最小二乘法(nonlinear iterative partial least squires，NIPALS)。这一方法以所需计算机内存少、易于实现著称。

以对 m 个变量的 n 次观测值组成一个矩阵为例

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

假定 $\boldsymbol{X}_{n\times m}$的秩为 r[$r<\min(n,m)$]，可以将 $\boldsymbol{X}$ 写成 r 个秩为 1 的矩阵之和

$$\boldsymbol{X}=\boldsymbol{Z}_1+\boldsymbol{Z}_2+\cdots+\boldsymbol{Z}_h+\cdots+\boldsymbol{Z}_r$$

这些秩为 1 的矩阵 $\boldsymbol{Z}_h$，可以表示为两个向量的外积[其中向量之一 $\boldsymbol{t}_h$ 称为得分向量(score)，维数为 n；另一向量 $\boldsymbol{p}_h$ 称为载荷向量(loading)，维数为 m]，即 $\boldsymbol{Z}_h=\boldsymbol{t}_h\boldsymbol{p}_h'$，维数为 $n\times m$，与 $\boldsymbol{X}$ 的相同。因此，上式可写为

$$\boldsymbol{X}=\boldsymbol{t}_1\boldsymbol{p}_1'+\boldsymbol{t}_2\boldsymbol{p}_2'+\cdots+\boldsymbol{t}_h\boldsymbol{p}_h'+\cdots+\boldsymbol{t}_r\boldsymbol{p}_r'$$

其矩阵形式表示为

$$\boldsymbol{X}=\boldsymbol{TP}$$

其中，$\boldsymbol{T}$ 和 $\boldsymbol{P}$ 分别由 $\boldsymbol{t}$ 和 $\boldsymbol{p}$ 组成。

通常，主成分分析中，人们所关心的是投影操作。通过一种操作使 $\boldsymbol{X}_{n\times m}$ 向一维空间投影，使得它的每一列用一个标量表示；通过另一种操作使 $\boldsymbol{X}_{n\times m}$ 向另一维空间投影，使得它的每一行用一个标量表示。这些操作具有很简单的特点。最常用的操作是 NIPALS 方法。NIPALS 方法并不能一次计算出所有主成分。它首先从 $\boldsymbol{X}_{n\times m}$ 计算出 $\boldsymbol{t}_1$ 和 $\boldsymbol{p}_1$，然后，从 $\boldsymbol{X}_{n\times m}$ 减去 $\boldsymbol{t}_1$ 和 $\boldsymbol{p}_1'$ 的外积 $\boldsymbol{t}_1\boldsymbol{p}_1'$，得到残差矩阵 $\boldsymbol{E}_1$。再用 $\boldsymbol{E}_1$ 计算 $\boldsymbol{t}_2$ 和 $\boldsymbol{p}_2$，即

$$\boldsymbol{E}_1 = \boldsymbol{X} - \boldsymbol{t}_1\boldsymbol{p}_1', \quad \boldsymbol{E}_2 = \boldsymbol{E} - \boldsymbol{t}_2\boldsymbol{p}_2', \quad \cdots, \quad \boldsymbol{E}_h = \boldsymbol{E}_{h-1} - \boldsymbol{t}_h\boldsymbol{p}_h'$$

NIPALS 代数如下：

1) 从 $\boldsymbol{X}_{m\times n}$ 中任取一向量 $\boldsymbol{x}_j$，将其赋值给 $\boldsymbol{t}_h$，$\boldsymbol{t}_h = \boldsymbol{x}_j$；

2) 计算 $\boldsymbol{p}_h'$，即 $\boldsymbol{p}_h' = \boldsymbol{t}_h'\boldsymbol{X}/\boldsymbol{t}_h'\boldsymbol{t}_h$；

3) 将 $\boldsymbol{p}_h'$ 归一化，即 $\boldsymbol{p}_{h新}' = \boldsymbol{p}_{h旧}'/\|\boldsymbol{p}_{h旧}'\|$；

4) 计算 $\boldsymbol{t}_h$，即 $\boldsymbol{t}_h = \boldsymbol{X}\boldsymbol{p}_h/\boldsymbol{p}_h'\boldsymbol{p}_h$；

5) $\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{t}_h\boldsymbol{p}_h'$；

6) 比较步骤 4)和 2)中的 $\boldsymbol{t}_h$，如果迭代收敛，停止计算，否则转至步骤 2)继续迭代。

可以看出，第二步中的 $\boldsymbol{t}_h'\boldsymbol{t}_h$，第三步中的 $\|\boldsymbol{p}_{h旧}'\|$ 和第四步中的 $\boldsymbol{p}_h'\boldsymbol{p}_h$ 都是标量(即常数)，即不存在求逆运算。

总之，一个秩为 r 的矩 $\boldsymbol{X}$ 可以分解为 r 个秩为 1 的矩阵之和；这些秩为 1 的矩阵之和是得分向量和载荷向量的乘积。

另一种计算方法是协方差矩阵分解方法：

1) 计算协方差矩阵

$$\boldsymbol{Z} = \boldsymbol{X}'\boldsymbol{X}$$

2) 得分向量初始赋值

$$\boldsymbol{T}_i = 0.1$$

3) 计算新得分

$$\boldsymbol{T}_i = \boldsymbol{Z}\boldsymbol{T}_i'$$

4) 计算特征值

$$\lambda_{i,i} = \left(\sum \boldsymbol{T}_i^2\right)^{1/2}$$

5) 得分向量归一化

$$\boldsymbol{T}_i = \boldsymbol{T}_i/\lambda_{i,i}$$

6) 检验 $\boldsymbol{T}$ 是否收敛，如果不收敛，返回步骤 3)，否则执行步骤 7)；

7) 如果 $i = f$，执行步骤 9)，否则为下一个特征向量计算协方差矩阵

$$\boldsymbol{Z} = \boldsymbol{Z} - (\boldsymbol{T}_i\boldsymbol{T}_i')\lambda_{i,i}$$

8) $i = i+1$，返回步骤 3)；

9）计算所有的特征向量

$$\boldsymbol{P}=\boldsymbol{T}'\boldsymbol{X}$$

2.4　基于主成分分析的多元校正方法[10]

在化学计量学中最常用的多变量校正方法包括多元线性回归、主成分回归、偏最小二乘等。

2.4.1　多元线性回归法

多元线性回归方法，又称逆最小二乘法或 P 矩阵法。当测得了一个未知样的响应矢量和已知各待测组分的单位浓度响应值时，可应用多元线性回归方法去预测未知样中各待测组分的浓度。当用多个（混合）标准溶液去建立校正模型时，利用多样本的多元线性回归方法可求取系数矩阵，然后再求取多个未知样的各待测组分的浓度值。下面先简要介绍多元线性回归法。

若自变量为 $x_j(j=1,2,\cdots,m)$，一个因变量为 y，用多元线性回归方法，建立因变量 y 和 x_j 之间的线性关系。其数学模型为

$$\boldsymbol{y}=b_1x_1+b_2x_2+\cdots+b_mx_m+e$$

$$\boldsymbol{y}=\sum_{j=1}^{m}b_jx_j+e$$

$$\boldsymbol{y}=\boldsymbol{x}'\boldsymbol{b}+e$$

其中

$$\boldsymbol{x}'=[x_1,x_2,\cdots,x_m],\quad \boldsymbol{b}=[b_1,b_2,\cdots,b_m]'$$

如果有 n 个样本，则

$$\boldsymbol{y}=[y_1,y_2,\cdots,y_n]'$$

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}+e$$

其中

$$\boldsymbol{X}=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1m}\\ x_{21} & x_{22} & \cdots & x_{2m}\\ \vdots & \vdots & & \vdots\\ x_{n1} & x_{n2} & \cdots & x_{nm}\end{bmatrix}$$

对于该数学模型的解有 3 种情况。

1）当 $m>n$，即变量数多于样本数，$\boldsymbol{b}$ 有无穷多个解。

2）当 $m=n$ 时，如果 $\boldsymbol{X}$ 满秩，$\boldsymbol{b}$ 有惟一的解。

$$\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}=0$$

3）当 $m<n$ 时，变量数小于样本数，得不到精确的解。

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}$$

其最小二乘解为

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

如果有 k 个因变量，即 $\boldsymbol{Y}$ 矩阵中有 k 列，则数学模型为

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$$

其最小二乘解为

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

$$\boldsymbol{B} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ b_{21} & b_{22} & \cdots & b_{2k} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mk} \end{bmatrix}$$

多元线性回归表面上看似乎是，只要知道混合物中某些组分的浓度或性质，就可以建立复杂体系的校正模型。惟一的要求就是选择对应于被测组分数据向量（如在某些波长处的光谱吸收值向量）。但多元线性回归存在许多弱点。如果 $\boldsymbol{X}$ 是光谱矩阵，光谱的强度在某些通道（波长）处往往成比例，这样会产生共线性问题。所谓共线性问题是指 $\boldsymbol{X}$ 矩阵是不满秩，即 $\boldsymbol{X}$ 中至少有一列或一行是可用其他几列或几行的线性组合表示出来。共线性问题可导致行列式值为零或接近于零，因此就无法用一般逆求矩阵的逆。

再者多元线性回归使用 $\boldsymbol{X}$ 矩阵建立模型，并没有考虑 $\boldsymbol{X}$ 矩阵中的信息是否与真实模型相关。如果使用的变量包括了噪声，就会导致过度拟合情况，影响模型的预测能力。

上述缺点限制了多元线性回归方法不能使用太多的向量参与回归。因此，多元线性回归的重要任务就是如何选择参加回归的变量，逐步多元线性回归就是为解决这个问题而发展的方法。但对于实际问题来说，并非容易，如一张光谱包含2000多个变量（波长通道），其筛选工作量是巨大的。实际工作中，往往靠人工凭化学知识进行选择，譬如根据苯的紫外光谱的最大吸收位置在254nm处来选择苯的回归变量。

一般来说，要建立一个可靠的多元线性回归模型是需要较多样品的，而收集样品和测量数据的工作是比较艰巨的。

2.4.2 主成分回归法

主成分回归法（PCR）能够有效地解决多元线性回归中遇到的共线性问题、变量数使用限制问题和在一定程度上解决了噪声滤除问题。当用于多元校正时，由建立校正模型和预测两个阶段所组成。其运算也分直接计算和利用NIPALS原理两种。为了便于理解，下面以光谱分析为例说明利用NIPALS原理的主成分回

归法。

在实际工作中遇到的光谱，构成光谱的因素很多，如样品的成分、成分之间的相互作用、光谱仪的影响（检测器噪声、环境对基线的影响等）和样品的前处理等。尽管有许多因素对光谱都有贡献，但是总会有一定数目的独立变量存在于光谱中。理想情况下，希望校正集光谱中的最大变化最好是被测样品的组成或性质的变化引起。模型的建立是根据光谱的这种变化，而不是根据光谱的绝对强度。实际工作中，光谱的变化是由上述多种因素引起，也可以将光谱看成是这些因素各自产生的各种光谱的加和。每种光谱都可以看成是自己的“纯光谱”乘上一个权重（标量）得到的。换句话说，将所有这些“纯光谱”乘上其相应的权重后再加和，就能重建样品的原始光谱图。

在实际计算中，利用数学方法对矩阵分解，只能得到特征向量（eigenvector）和特征值。在化学计量学中称之为载荷向量，或因子，或主成分（principal component）。这些权重则称为得分。这些特征向量是相互独立的，即互为正交，如果使用它们进行回归，就不会产生共线性问题。因为各个特征向量分别代表了原始光谱中包含的不同因素的贡献，如果我们能有效地选择仅与被测组分或性质有关的特征向量参加回归运算，就排除了光谱中包含的噪声对模型的不利影响。使用特征向量回归，就会克服多元线性回归中必须限制光谱通道数的问题。它可以使用全谱，也可以使用原始光谱的一部分。这些载荷向量和得分向量的计算方法就是在 2.3 节中介绍的主成分分析（PCA）。

主成分回归过程分为两步：主成分分析和多元线性回归。

这里自变量 $\boldsymbol{X}$ 和因变量 $\boldsymbol{Y}$ 与在多元线性回归中的规定相同，即

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \qquad \boldsymbol{Y}=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1k} \\ y_{21} & y_{22} & \cdots & y_{2k} \\ \vdots & \vdots & & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nk} \end{bmatrix}$$

首先，使用 PCA 对 $\boldsymbol{X}$ 矩阵运算求出载荷矩阵 $\boldsymbol{P}$ 和得分矩阵 $\boldsymbol{T}$

$$\boldsymbol{T}=\boldsymbol{XP}$$

$\boldsymbol{T}$ 矩阵的维数可以与矩阵 $\boldsymbol{X}$ 的相同，如果使用整个 $\boldsymbol{T}$ 矩阵参加回归，这样得到的结果和多元线性回归的结果没有多大区别，因为所谓载荷（主成分）就是新变量，它是原变量的线性组合。第一主成分反映了原变量的最大方差变化，第二主成分次之，第三主成分更次之，依此类推，即前边的主成分包含了 $\boldsymbol{X}$ 矩阵的绝大部分有用信息，而后边的主成分则往往与噪声和干扰影响因素有关，因此主成分分析中参与回归的 $\boldsymbol{T}$ 是选取前边少数主成分组成的矩阵，在维数上远小于 $\boldsymbol{X}$。

将降维后的 $\boldsymbol{T}$ 与因变量 $\boldsymbol{Y}$ 矩阵进行多元线性回归

$$\boldsymbol{Y}=\boldsymbol{TB}+\boldsymbol{E}$$

其中最小二乘解为

$$\boldsymbol{B} = (\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{Y}$$

预测分析

$$\boldsymbol{Y}_{未知} = \boldsymbol{T}_{未知}\boldsymbol{B} = \boldsymbol{X}_{未知}\boldsymbol{P}\boldsymbol{B}$$

主成分回归通过对参与回归的主成分的合理选取可以去掉噪声。由于 $\boldsymbol{T}$ 的各列互相正交,解决了多元线性回归中的共线性问题。在主成分回归中,仅对 $\boldsymbol{X}$ 矩阵进行了降维处理或矩阵分解,而对 $\boldsymbol{Y}$ 矩阵的噪声并没有考虑。通过主成分分析得到一系列主成分,但某些主成分和因变量 $\boldsymbol{Y}$ 之间不一定有相关关系,因此这样的回归结果就不一定合理。

(1) PCR 的优点

1) 可以使用整体量测数据(原始全谱或部分谱数据)。能充分利用数据信息。使用更多的数据则能利用数据的平均效应,增加模型抗噪声干扰的能力。

2) 通过主成分选择,可有效地滤除噪声。

3) 解决了共线性问题。

4) 适用于部分复杂分析体系,只要标样包含了可能存在的组分,就可以预测组分。

(2) PCR 的缺点

1) 计算速度比多元线性回归慢。

2) 模型优化需要 PCA,对模型的理解不如多元线性回归直观,较难理解。

3) 并不能保证将参与回归的主成分一定与被测组分或性质相关。

4) 不适用于存在未知干扰组分的复杂分析体系。

2.4.3 偏最小二乘法

2.4.3.1 原理

在 PCR 中,只对 $\boldsymbol{X}$ 矩阵做了分解,消除 $\boldsymbol{X}$ 矩阵中无用的信息。同样,$\boldsymbol{Y}$ 矩阵中也包含了无用信息,也应做同样的处理,PLS 就是基于这种思想的回归方法。它基于 NIPALS 思路,同时对 $\boldsymbol{X}$ 矩阵和 $\boldsymbol{Y}$ 矩阵逐个主组分提取有用信息,直到一定主成分时建立线性回归模型。

PLS 的第一步,作矩阵分解,其模型为

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P} + \boldsymbol{E}$$

$$\boldsymbol{Y} = \boldsymbol{u}\boldsymbol{Q} + \boldsymbol{F}$$

式中:$\boldsymbol{T},\boldsymbol{U}$——$\boldsymbol{X}$ 矩阵和 $\boldsymbol{Y}$ 矩阵的得分矩阵;

$\boldsymbol{P},\boldsymbol{Q}$——$\boldsymbol{X}$ 矩阵和 $\boldsymbol{Y}$ 矩阵的载荷(即主成分)矩阵;

$\boldsymbol{E},\boldsymbol{F}$——用 PLS 模型拟合 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 时所引进的误差。

PLS 的第二步,将 $\boldsymbol{T}$ 和 $\boldsymbol{U}$ 作线性回归。

$\boldsymbol{B}$ 为关联系数矩阵

$$\boldsymbol{U} = \boldsymbol{TB}$$

$$\boldsymbol{B} = \boldsymbol{TU}(\boldsymbol{T}'\boldsymbol{T})^{-1}$$

在预测时，由未知样品的矩阵 $\boldsymbol{X}_{未知}$ 和校正得到的 $\boldsymbol{P}_{校正}$ 求出未知样品 $\boldsymbol{X}$ 矩阵的 $\boldsymbol{T}_{未知}$。然后得到

$$\boldsymbol{Y}_{未知} = \boldsymbol{T}_{未知}\boldsymbol{BQ}$$

实际上，PLS 计算并非如此，PLS 把矩阵分解和回归并为一步，即 $\boldsymbol{X}$ 矩阵和 $\boldsymbol{Y}$ 矩阵的分解是同时进行的，并且将 $\boldsymbol{Y}$ 信息引入到 $\boldsymbol{X}$ 矩阵分解过程中，在每计算一个新主成分之前，将 $\boldsymbol{X}$ 得分和 $\boldsymbol{Y}$ 得分进行交换，使得到 $\boldsymbol{X}$ 主成分直接与 $\boldsymbol{Y}$ 关联。这就不同于 PCR，PCR 在 $\boldsymbol{X}$ 矩阵分解时并不考虑 $\boldsymbol{Y}$ 的影响。

PLS 又分为 PLS1 和 PLS2。虽然方法差别不大，但所得到的结果却有很大差异。所谓 PLS2 和 PCR 相似，在校正过程中，$\boldsymbol{X}$ 矩阵分解只给出一个 $\boldsymbol{T}$ 矩阵和一个 $\boldsymbol{P}$ 矩阵，显然，这样得到的 $\boldsymbol{T}$ 和 $\boldsymbol{P}$ 是对 $\boldsymbol{Y}$ 中的个别向量并不是最优化的，在预测时，对于复杂体系，会降低结果精度；在 PLS1 中，校正得到的 $\boldsymbol{T}$ 和 $\boldsymbol{P}$ 是对 $\boldsymbol{Y}$ 中的个别向量进行优化的，换句话说，对应于 $\boldsymbol{Y}$ 矩阵中不同的向量，其 $\boldsymbol{T}$ 和 $\boldsymbol{P}$ 矩阵不同。当构成训练集的样品中组分浓度变化上相差很大时，如一个组分浓度范围为 60%～70%，而另一个组分为 0.2%～0.8%，PLS1 预测结果普遍优于其他方法。其缺点是在校正时，对每个被分析组分都要计算一套主成分矩阵和得分矩阵，需要的校正计算时间比 PCR 和 PLS2 长。如果校正集比较大时，PLS1 校正时间长的问题将比较突出，但对预测的影响并不大。由于计算机计算速度越来越快，这些问题已不再成为问题了。

2.4.3.2　算法

(1) 算法推导

对于 $\boldsymbol{X}$ 矩阵：

1) 将 $\boldsymbol{X}$ 矩阵的任意一列 x_j 赋值给 t_0，即 $t_0 = x_j$；

2) $p' = t'\boldsymbol{X}/t't$；

3) $p'_{新} = p'_{旧}/\| p_{旧} \|$；

4) $t = \boldsymbol{X}p/p'p$；

5) 比较步骤 4)和步骤 2)中的 t，如果收敛，停止迭代，否则转到步骤 2)继续循环。

对于 $\boldsymbol{Y}$ 矩阵：

1) $u_0 = y_i$；

2) $q' = u'\boldsymbol{Y}/u'u$；

3) $q'_{新} = q'_{旧}/\| q'_{旧} \|$；

4) $u=\boldsymbol{Y}q/q'q$。

比较步骤 4)和步骤 2)中的 u,如果收敛,停止迭代,否则,转到步骤 2)继续循环。

上述计算步骤相互独立求出的 t 和 u,仍与 PCR 没有质的区别。为了建立两者内在的相关性,将得分 t 和 u 在步骤 2)中的位置相交换:

1) $u_0=y_i$;

2) $p'=u'\boldsymbol{X}/u'u$〔这里以 u 代替 t〕;

3) $p'_{新}=p'_{旧}/\|p'_{旧}\|$;

4) $t=\boldsymbol{X}p/p'p$;

5) $q'=t'\boldsymbol{Y}/t't$〔这里以 t 代替 u〕;

6) $q'_{新}=q'_{旧}/\|q'_{旧}\|$;

7) $u=\boldsymbol{Y}q/q'q$;

8) 将步骤 4)中的 t 和前一次迭代的 t 比较,如果收敛,转到步骤 6),否则到步骤 2)继续循环;

由于交叉分解得到的 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 主成分(t)并不互相正交,因此需要在步骤 2)中以权重(w)代替 p,并在收敛后,再加入。

9) $p'=t'\boldsymbol{X}/t't$。t 的相互正交并非绝对必要,但当它与主成分回归比较时,t 正交的条件还是必要的。当预测时,w 也需做标准化处理,否则将引入误差。

10) $p'_{新}=p'_{旧}/\|p'_{旧}\|$;

11) $t=\boldsymbol{X}p/p'p$,然后,t 可用于内部相关;

12) $b_f=u_f t/t'_f t_f$;

13) $u_f=b_f t_f$,$f=1,2,\cdots,n$,其中 f 为主成分序数。

计算残差:

1) $E_f=E_{f-1}-t_f p'_f$;

2) 令 $\boldsymbol{X}=E_f$;

3) $F_f=F_{f-1}-u_f q'_f$;

4) 令 $\boldsymbol{Y}=F_f$。

(2) PLS 算法

1) 校正部分。

① 在计算前首先将数据标准化。分别将矩阵 $\boldsymbol{X}_{n\times m}$ 和 $\boldsymbol{Y}_{n\times k}$ 中各列中心化及规格化,即列中各元素减去该列平均值,所得差除以该列标准方差

$$S_{x,j}=\sqrt{\left[\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2/(n-1)\right]}$$

$$S_{y,j}=\sqrt{\left[\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2/(n-1)\right]}$$

$$x_{ij} = (x_{ij} - \bar{x}_j)/S_j, \quad i = 1,2,\cdots,n; \quad j = 1,2,\cdots,m$$
$$y_{ih} = (y_{ih} - \bar{y}_h)/S_h, \quad i = 1,2,\cdots,n; \quad h = 1,2,\cdots,k$$

② 迭代次数变量赋值(f=1,2,…,d)。

③ 初始迭代变量赋值。将 Y 的任意一列赋值给初始的 u。

④ 计算 $\boldsymbol{X}$ 矩阵的权重变量

$$w' = u'\boldsymbol{X}/u'u$$

⑤ 权重变量归一化

$$w'_{新} = w'_{旧}/\| w'_{旧} \|$$

⑥ 计算 $\boldsymbol{X}$ 矩阵的 t 变量

$$t = Xw/w'w$$

⑦ 计算 $\boldsymbol{Y}$ 矩阵的 q 变量

$$q' = t'\boldsymbol{Y}/t't$$

⑧ q 变量归一化

$$q'_{新} = q'_{旧}/\| q'_{旧} \|$$

⑨ 计算 $\boldsymbol{Y}$ 矩阵的 u 变量

$$u = \boldsymbol{Y}q/q'q$$

⑩ 收敛检验。检验 t 与前一轮的 t,如果收敛,转到步骤⑪,否则转到步骤④继续迭代。

⑪ 计算 $\boldsymbol{X}$ 矩阵的 p 变量

$$p' = t'\boldsymbol{X}/t't$$

⑫ p 变量归一化

$$p'_{新} = p'_{旧}/\| p'_{旧} \|$$

⑬ t 变量的正交化

$$t_{新} = t_{旧} \| p_{旧} \|$$

⑭ w 变量标准化

$$w'_{新} = w'_{旧} \| w'_{旧} \|$$

⑮ 计算回归系数

$$b = u't/t't$$

⑯ 计算残差矩阵及对 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 矩阵重新赋值

$$E_f = E_{f-1} - t_f p'_f$$
$$令\ \boldsymbol{X} = E_f$$
$$F_f = F_{f-1} - u_f q'_f$$
$$令\ \boldsymbol{Y} = F_f$$

⑰ 将计算得到的变量 t,p,u,q,b 保存起来,供预测使用。返回步骤③计算下一个主成分。

2）预测部分。

① 按校正部分的标准化处理方法将 $\boldsymbol{X}$ 矩阵标准化（特别注意的是，这里应使用校正集的均值和方差进行标准化，而不是预测集的均值和方差）。

② 逐维迭代，$f=1,2,\cdots,d$。

③ Y 的初值，当 $f=0$ 时（$\boldsymbol{Y}$ 的均值是校正集的），有

$$\boldsymbol{Y}=\bar{y}$$

④ 当 $f=f+1$ 时（这里使用了 W,b,q 都是校正集的），有

$$t_f = XW'_f$$

$$y = y + b_f\, t_f\, q'_f$$

⑤ 计算 $\boldsymbol{X}$ 矩阵残差（这里使用的 p 是校正集的）

$$x = x - t_f p'_f$$

⑥ 如果 $f<d$，转到步骤②继续迭代。

2.4.3.3 特点

（1）PLS 的优点

1）可以使用全谱或部分谱数据。

2）数据矩阵分解和回归交互结合为一步，得到的特征值向量直接与被测组分或性质相关，而不是与数据矩阵中变化最大的变量相关。

3）如果选择的校正集具有代表性，PLS 模型更稳健。

4）可以使用于校正阶段与预测阶段的模型不存在系统误差的部分复杂分析体系。

（2）PLS 的缺点

1）计算速度相对较慢。

2）模型建立过程复杂，较抽象，较难理解。

3）与 PCR 相似，不适用于存在未知干扰组分的复杂分析体系的多组分同时定量分析。

2.5 主成分数的确定[10]

主成分数确定问题存在于矩阵分解、主成分回归、PLS 回归等应用中。在基于直接特征分解或奇异值分解的各类应用中，主成分数确定是一个实际存在但不突出的问题，因像 MATLAB 语言这样的软件，具有性能很好的子程序用于这样的计算。但在基于 NIPALS 原理编写的 PCR 和 PLS 程序中，考虑到内存及计算性能，需对主成分数进行预先估计。

2.5.1　量测矩阵的主成分数确定

量测矩阵的主成分数估计或确定就是线性代数中的秩估计。数学中的矩阵一般不具有物理意义，而现代科学仪器测量得到的响应数值矩阵既有大小又有测量单位，如 HPLC-DAD 数据、EEM 数据等。对一个复杂分析体系而言，体系中到底存在多少组分或化学成分是一个首先要考虑的问题。由于各类测量误差的存在，确定一个量测矩阵的主成分数显然比线性代数中的矩阵秩估计要复杂的多。但线性代数中的矩阵秩估计仍然是量测矩阵的主成分数确定的主要方法。对于一个存在多个性质极其相似组分的复杂分析体系而言，确定是否还存在未知的组分就是一个分析科学的难题。针对量测体系实际，结合线性代数中的秩估计算法是一个较好的做法。值得注意的是，既不能将误差当作实际存在的有物理意义的成分，也不能少估计实际存在的有物理意义的成分数。在 MATLAB 中，一般用 rank(X, Tol)可得到矩阵 X 大于 Tol 值的奇异值。

2.5.2　主成分回归中的主成分数确定

详见 2.4.2 节。

2.5.3　PLS 回归中的主成分数确定

使用 PCR 和 PLS 方法建立校正模型，其中最困难的问题之一就是如何确立建立模型所使用的主成分数目。在计算的多个主成分中，第一主成分最重要，随主成分数增加，重要程度依次降低，直到后来的许多主成分反映的是噪声信息。因此，前面的主成分在建立模型时比后面的主成分更有用。如果建立模型时使用的主成分数过少，就不能反映未知样品被测组分产生的量测数据(如光谱)变化，其模型预测准确度就会降低，这种情况称为欠拟合(underfit)。如果使用过多的主成分建立模型，就会将一些代表噪声的主成分加到模型中，使模型的预测能力下降这种情况称之为过拟合(overfit)。因此，合理确定参加建立模型的主成分数是充分利用光谱信息和滤除噪声的有效方法之一。

很遗憾的是，现在尚无明确的方法告诉我们，到底使用多少主成分才能保证既能避免不充分拟合又能避免过度拟合情况的出现。尽管如此，仍有许多方法可以帮助我们寻找正确的主成分数目。其中，最常用的是预测残差平方和(prediction residual error sum of squares，PRESS)。PRESS 是这样计算的：使用一定数目的主成分建立一个模型，然后用这个模型对参加建模的每个样品进行预测，求出每个样品的预测值和已知值的差，按照下式计算 PRESS

$$\mathrm{PRESS} = \sum_{i=1}^{n} \sum_{j=1}^{d} \left(y_{p,ij} - y_{ij} \right)^2$$