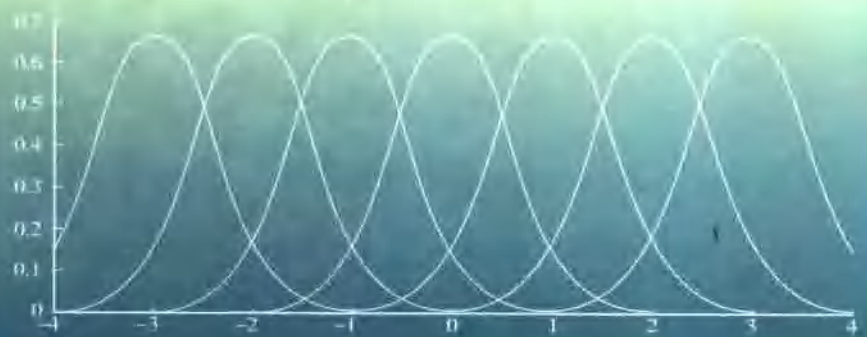
偏最小二乘回归的线性与非线性方法

Partial Least-Squares Regression —Linear and Nonlinear Methods

王惠文 吴载斌 孟洁 著



国防工业出版社
National Defense Industry Press

偏最小二乘回归的线性与非线性方法

Partial Least - Squares Regression ——Linear and Nonlinear Methods

王惠文　吴载斌　孟洁　著

国防工业出版社
·北京·

图书在版编目(CIP)数据

偏最小二乘回归的线性与非线性方法/王惠文，吴载斌，孟洁著. 北京：国防工业出版社，2006.9

ISBN 7-118-04496-2

Ⅰ.偏… Ⅱ.①王…②吴…③孟… Ⅲ.①回归－最小二乘法－线性－研究②回归－最小二乘法－非线性－研究 Ⅳ.O212.1

中国版本图书馆 CIP 数据核字(2006)第 031069 号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号 邮政编码 100044)

京南印刷厂印刷

新华书店经售

*

开本 850×1168 1/32 **印张** 10¾ **字数** 265 千字

2006 年 9 月第 1 版第 1 次印刷 **印数** 1—2000 册 **定价** 38.00 元

(本书如有印装错误，我社负责调换)

国防书店：(010)68428422 发行邮购：(010)68414474

发行传真：(010)68411535 发行业务：(010)68472764

致　读　者

国防科技事业已经取得了举世瞩目的成就。国防科技图书承担着记载和弘扬这些成就，积累和传播科技知识的使命。在改革开放的新形势下，原国防科工委率先设立出版基金，扶持出版科技图书，这是一项具有深远意义的创举。此举势必促使国防科技图书的出版随着国防科技事业的发展更加兴旺。

设立出版基金是一件新生事物，是对出版工作的一项改革。因而，评审工作需要不断地摸索、认真地总结和及时地改进，这样，才能使有限的基金发挥出巨大的效能。评审工作更需要国防科技和武器装备建设战线广大科技工作者、专家、教授，以及社会各界朋友的热情支持。

让我们携起手来，为祖国昌盛、科技腾飞、出版繁荣而共同奋斗！

国防科技图书出版基金
评审委员会

国防科技图书出版基金
第四届评审委员会组成人员

名誉主任委员	陈达植				
顾　　问	黄　宁				
主 任 委 员	刘成海				
副主任委员	王　峰	张涵信	张又栋		
秘　书　长	张又栋				
副 秘 书 长	彭华良	蔡　镭			
委　　员	于景元	王小谟	甘茂治	冯允成	
（按姓名笔画排序）	刘世参	杨星豪	李德毅	吴有生	
	何新贵	佟玉民	宋家树	张立同	
	张鸿元	陈火旺	侯正明	常显奇	
	崔尔杰	韩祖南	舒长胜		

序

偏最小二乘回归(Partial Least－Squares Regression,PLS 回归)是一种先进的多元分析方法。这一方法是伍德(S. Wold)和阿巴诺(C. Albano)等人于 1983 年首次提出的,用以解决化学样本分析中存在的变量多重相关,以及解释变量多于样本点等实际问题。由于 PLS 回归能解决许多以往用普通多元回归方法无法解决的问题,因而得到了有关研究人员的重视,在线性与非线性的 PLS 回归的理论及方法上发展迅速,其实际应用也不断扩展,涉及化学、机械、工业、生物、地质、医学、药物学、社会学以及经济学等领域。

王惠文教授于 1994 年开始研究并应用 PLS 回归,是我国最早引入这一方法的学者之一。由于她具有深厚的统计学功底,又勤奋好学,因而很快就掌握了这一先进的方法,并用它来研究一些复杂问题,取得了许多有价值的成果。她和她的合作者们曾应用 PLS 回归方法对中国城市的经济发展状况进行分析,研究在各类城市中,经济变量之间的相互作用方式,并运用一种特殊的 PLS 通径分析模型,建立了城市经济发展水平的评估指数。他们还结合 PLS 回归方法,提出成分数据的一元回归和多元回归的建模策略,用于根据 GDP 和投资的三次产业结构来预测劳动就业需求的三次产业结构。在机械控制技术方面,他们应用 PLS 回归研究刀具磨损的状态建模中的影响因素,建立了刀具状态的估计预报模型。他们将 PLS 回归用于多状态定性变量的回归建模问题,从理论上证明了免耕法能较好地面定土壤,减少水分蒸发,

具有防止土壤沙化的各项优势。此外，他们还将 PLS 的 logistic 回归方法应用于鄱阳湖洪水灾害模式的分析，取得了有益的分析结果。

1999 年，在我兼任国家自然科学基金委员会管理科学部主任期间，王惠文教授向我介绍了他们的研究情况。我立即感到 PLS 回归是一种非常有用的工具，有可能用来解决非线性、非稳态、非参数、紧耦合的复杂问题。此后我一直非常关注他们在该领域的研究进展。我将我在复杂数据分析方面的一些构思提供给他们参考，他们也参与了我主持的股市及期货方面的研究项目，并应用复杂数据分析方面的优异成果支持了我的研究。因此当她求我为其新著作序时，我也就义不容辞地欣然命笔。

王惠文教授曾于 1999 年撰写了《偏最小二乘回归方法及其应用》一书，该书以独到的理论视角，采用理论与实践相结合的方式，详细介绍了 PLS 回归的理论方法和分析技术，受到了学术界及读者的好评，并被广泛引用。在她与吴载斌、孟洁合作的新著——《偏最小二乘回归的线性与非线性方法》中，从理论及实践 2 个方面都有了很大的扩充。本书更加全面地体现了 PLS 回归近年来的理论与应用发展趋势，详细介绍了多种新的线性与非线性 PLS 回归技术，并反映了作者们的最新工作成果。特别是新增了 PLS 回归最优子空间的辨识和变量筛选方法、PLS 回归的非线性方法、PLS 的通径分析方法及其拓展等方面的内容，本书还在国内首次介绍了伍德和他的合作者所开发的在 Windows 下运行的 SIMCA－P 数据分析软件及应用实例。

本书从实用出发，采用理论与实践相结合的方式，介绍了 PLS 回归线性与非线性方法的理论方法和分析技术。书中特别着重讲述了各种 PLS 回归线性与非线性方法的应用案例分析，这些案例绝大部分源于作者与其合作者在经济管理与工程技术领域应用 PLS 回归技术的研究成果。我相信本书的出版将有助

于工程技术人员和经济管理工作者更全面地了解和掌握 PLS 回归线性与非线性方法的理论基础、方法特色、应用技巧和发展前景，并应用这一先进的工具来有效地解决他们面临的实际问题。

成思危

2006 年 3 月 6 日

前　言

偏最小二乘回归(Partial Least-Squares Regression，PLS 回归)是一种先进的多元分析方法。它于 1983 年由伍德(S. Wold)和阿巴诺(C. Albano)等人首次提出，主要用来解决多元回归分析中的变量多重相关性或解释变量多于样本点等实际问题。由于偏最小二乘回归解决了这些以往用普通多元回归方法无法解决的难题，因此该方法的理论研究进展非常迅速，而其应用领域已经从最初的化工领域快速扩展到机械、生物、地质、医学、社会学以及经济学等领域。

偏最小二乘回归在 20 世纪 90 年代被介绍到中国。在国防科技图书出版基金的支持下，1999 年本书作者之一出版了专著《偏最小二乘回归方法及其应用》。该书曾力图以独到的理论视角，采用理论与实践相结合的方式，系统介绍偏最小二乘回归方法及其辅助分析技术。然而，由于当时理论发展的局限性，书中的内容还主要集中在偏最小二乘回归的普通线性模型的研究，而有关更为复杂的偏最小二乘线性方法、偏最小二乘回归的非线性模型以及变量筛选和模型解释等问题都还没能涉及。近年来，偏最小二乘回归在理论方法方面取得了许多突破性的进展，因而更进一步拓宽了偏最小二乘回归的应用范畴。为了进一步介绍偏最小二乘回归的理论与应用发展趋势，同时反映作者新近的工作进展，本书将在上一本论著的基础上，较为深入地讨论偏最小二乘回归中的以下一些新的理论课题。

1)偏最小二乘的通径分析方法和递阶模型

近年来，一些更为复杂的偏最小二乘回归的线性方法在欧洲取得了引人注目的进展。其中最为典型的是**偏最小二乘通径分析**

方法。该方法在变量多重相关条件下，为研究多组变量集合间的相互联系方式提供了有效的技术途径。偏最小二乘通径分析方法可拓展的空间也非常广泛。2002 年成思危教授在瑞士访问国际管理发展研究院(IMD)期间，曾针对他们目前在构造各国竞争力比较指标的方法中存在的缺陷，提出在建立综合评价指数过程中应注意多组变量集合中的多重相关问题。而该问题正是通过采用一种特殊的偏最小二乘通径分析，得到了满意的解决。另一个新的、应用性很强的偏最小二乘回归线性方法是**递阶偏最小二乘回归模型**。它通过分层次实施偏最小二乘回归，实现对多组变量集合的综合与建模目的。这个模型的构造简单，概念清晰，有效性强，特别适用于大规模变量集合的回归分析问题。此外，本书还应用偏最小二乘回归、偏最小二乘通径分析和递阶偏最小二乘回归模型，解决了一种非常特殊的数据类型——**成分数据的一元和多元回归建模**问题。该方法被应用于北京市的劳动就业需求预测，取得了满意的分析结果。

2)偏最小二乘回归的非线性方法

本书另一个新的重点内容是偏最小二乘回归的非线性方法。众所周知，自然界以及人类社会中的现象往往是复杂的、非线性的，只有较好地解决非线性建模问题，偏最小二乘回归方法的应用领域才能更加广阔。基于这个出发点，本书将在简要介绍一般非线性分析方法的基础上，重点讨论基**于样条变换的偏最小二乘回归方法**和基于**核函数变换的偏最小二乘回归方法**。此外，书中还将介绍**偏最小二乘 logistic 回归方法**及其在洪水灾害模式分析中的应用。

3)偏最小二乘回归理论的进一步探讨

经过偏最小二乘回归分析后，可以在原来高维自变量空间中找到一个低维的子空间，使之对因变量有更强的解释能力。然而长期以来，由于偏最小二乘回归的算法过程比较复杂，使得人们缺乏分析该低维子空间含义的技术方法，从而影响到对模型物理意义的基本认识。在本书中，作者采用正交变换方法，通过寻找新的

公共因素，从而达到更好地解释模型含义的目的。本书还利用非参数检验和再抽样(bootstrap)技术，给出偏最小二乘回归模型中的参数检验方法。该方法可以将一些解释作用很低的变量从模型中删除，从而得到一个更加合理有效的自变量集合。此外，吴喜之教授曾经举例说明，在有些情形下，用偏最小二乘方法提取的成分只对自变量集合有很强的综合能力，但与因变量的相关程度并不是最大的。针对这个问题，他对偏最小二乘回归方法做了进一步的改善，从而得到解释性更强的模型结果。本书详细地介绍了这一成果，进一步展示了偏最小二乘回归方法的理论发展空间。

4)SIMCA－P 软件介绍

伍德和他的合作者开发了在 Windows 下而运行的 SIMCA－P 数据分析软件，为偏最小二乘回归的计算和结果解释提供了有效的计算工具。为了使读者能够更加便捷地应用偏最小二乘回归方法，本书将简要介绍 SIMCA－P 软件的基本功能，并结合土壤风蚀的案例，演示使用 SIMCA－P 软件进行数据处理、建立偏最小二乘回归模型的全过程。通过这个案例，还说明了如何应用偏最小二乘回归方法，解决多状态定性变量的回归建模问题。

本书从实用的原则出发，采用理论与实践相结合的方式，重点介绍偏最小二乘回归线性与非线性的理论方法及其分析技术。本书还特别注重讲述各种方法的应用案例分析，案例研究贯穿在每一个重要的章节。而书中所选取的案例绝大部分取源于作者与其合作者在经济管理和工程技术领域应用偏最小二乘回归技术的研究成果，例如刀具磨损状态预报、地区发展水平综合评估、劳动就业需求预测、洪水灾害模式识别以及免耕法在防治土壤风蚀方而的效果评估等。通过这些实际课题研究的示范，可以帮助工程技术人员和经济管理工作者更全面地掌握偏最小二乘回归线性与非线性方法的基本功能、方法原理和应用技巧，使得偏最小二乘回归的线性和非线性方法真正成为他们手中的一个实用的工具。

本书在相关的研究工作中，曾受到国家杰出青年科学基金(编号 70125003)、国家自然科学基金重点项目(编号 70351010)、国家

自然科学基金创新研究群体科学基金项目(编号 70521001)、国家自然科学基金(编号 70371007)和北京市自然科学基金(编号 9052006)等项目的支持。作者在该领域的研究一直得到成思危教授的支持和帮助。成思危教授作为著名化工领域的专家以及系统科学、软科学与管理科学专家,十分注重推动复杂系统数据分析研究领域的前沿发展,并促进中外科学家建立密切的国际合作关系。他早在 90 年代初就开始关注偏最小二乘回归在化工科研课题中的应用问题,并对作者在该领域的研究给予了详细具体的指导。1997 年他首次提出偏最小二乘回归模型的含义辨识问题,由此构造了偏最小二乘回归模型的解释方法。2002 年,他再次提出解决多变量组评估中的变量多重相关问题,进一步促进了偏最小二乘回归前沿方法的应用进展。作者在长期的研究过程中,还受到德昂赫斯(M. Tenenhaus)和维兹(V. E. Vinzi)等国际著名专家的热情帮助。多年来,他们为作者的研究提供了大量的前沿资料和研究信息,并在理论探讨与应用项目中给予许多具体的指导。作者愿借此机会,向他们表示衷心的感谢。

在本书的研究内容和写作过程中,包含了很多同事和同学们的辛勤工作。除本书作者外,文中许多内容还直接来源于成思危教授、路明教授、吴喜之教授、刘强教授,以及朱韵华、付凌晖、李大鹏、黄薇等同学的研究工作。此外,龙文、张诤、赵安顺、马涛、张瑛、卢钰、王雅楠等同学也对本书的排版等工作给予了重要的帮助。中国人民大学的吴喜之教授非常鼓励和支持我们的研究工作。征求他的同意,本书将吴喜之教授关于偏最小二乘回归局限性的讨论以及改善方法作为书中的一节,提醒人们在相关领域的研究过程中,还有很多值得深入思考的问题。北京大学的耿直教授和北京航空航天大学陈懋章院士对本书的写作和出版给予了十分重要的支持。在本书出版之际,作者愿向他们表示最诚挚的谢意。

作者关于偏最小二乘回归专著的写作与出版,曾多次得到国防科技图书出版基金的支持,并承蒙评审委员会对本书的写作计划和部分初稿提出十分重要的建议。事实上,正是在这些重要的

支持和帮助下，作者才能够坚持长期跟踪与推进在偏最小二乘回归领域中的理论与应用的发展，不断取得新的研究成果。

作者还要特别感谢国防工业出版社的同志们为本书出版所付出的心血。没有他们的帮助和细致严谨的工作，本书是不会以这样的形式面世的。

由于作者的水平有限，书中难免存在缺点和错误，敬请读者批评指正。

作 者

2005 年 10 月

目　录