复杂社会经济行为建模与管理研究丛书

变量筛选、模型分类及自动化建模方法

王惠文 孟洁 著

科学出版社

复杂社会经济行为建模与管理研究丛书

变量筛选、模型分类及自动化建模方法

王惠文　孟　洁　著

科学出版社
北京

内 容 简 介

本书采用理论与实践相结合的方式，重点介绍了变量筛选、模型分类及自动化建模方法。在本书内容中，很多是直接来源于本书作者与其合作者在相关领域的研究工作，诸如关于变量多重相关对多元回归、主成分分析以及偏最小二乘回归等几种有代表性的多元分析模型的影响方式；基于Gram-Schmidt变换的无导师模型与有导师模型的变量筛选方法；多元分析建模的本征信息的分析，以及模型的动态预测方法、海量模型的分类方法和多元回归的自动化建模方法等。本书在写作过程中，特别重视其实用价值，对提及的技术方法都给出了仿真实验研究和应用案例研究，这些实际研究的示范可以帮助工程技术人员和管理工作者更全面地掌握相关理论方法的基本原理和应用技巧，使得变量筛选、模型分类及自动化建模方法成为他们手中的一个实用工具。

本书的读者对象为经济、管理、社会及工程等领域的科研人员，以及高等院校相关专业的研究生和高年级本科生。

图书在版编目（CIP）数据

变量筛选、模型分类及自动化建模方法 / 王惠文，孟洁著. —北京：科学出版社，2013

（复杂社会经济行为建模与管理研究丛书）

ISBN 978-7-03-037516-2

Ⅰ.①变… Ⅱ.①王…②孟… Ⅲ.①统计数据－统计分析（数学）－多元回归分析 Ⅳ.①O212.1

中国版本图书馆CIP数据核字（2013）第103659号

责任编辑：马 跃 / 责任校对：阴会宾
责任印制：徐晓晨 / 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

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2013年5月第 一 版 开本：720×1000 B5
2013年5月第一次印刷 印张：13 1/4
字数：250 000

定价：52.00元

（如有印装质量问题，我社负责调换）

丛书编委会成员

总 序

“基于行为的若干复杂社会经济系统建模与管理”是国家自然科学基金委员会管理科学部在我国高等院校资助的第三个创新研究群体项目，于2005年获得第一期资助（项目编号：70521001），2008年获得第二期资助（项目编号：70821061）。

北京航空航天大学经济管理学院拥有深厚的学术研究底蕴、扎实的学风和优秀的研究团队，其科研工作强调数理基础和学科交叉、理论与实践结合、定量与定性结合。学院已培养全国百篇优秀博士学位论文获得者3名、提名奖获得者2名、首届中金经济学/金融学优秀博士论文一等奖获得者1名，科研成果两次获得教育部自然科学一等奖和大量的省部级科技进步奖，每年在SCI/SSCI收录的国际学术刊物上发表论文40篇以上。

本创新群体致力于对复杂社会经济系统的运行机理和管理问题展开基于行为的建模理论研究，提出相应的优化管理措施，重点在三个既有特色又很典型的社会经济系统中展开实证研究，即交通系统的建模与管理、经济和金融系统的建模与管理、生产系统的建模与管理。虽然交通、经济金融和生产管理具有不同的行为发生和执行主体，但研究这些行为的方法是基本相同的。这是因为，这些系统虽然有一些共同且主要的特征，即个体行为都是理性的，但涌现出的整体表象确是难以预期的、甚至是不合理的。本创新群体对这些演变的原因、过程、结果以及控制方法展开深入研究，采用系统科学的理论和方法，发现内在的演化规律，设计有效的干预措施，这既有科学价值、又有实践指导意义。本创新群体在四个领域的主要工作包括：

(1)交通行为建模与分析。对交通需求进入路网的过程进行剖析，包括路径、交通工具和出发时刻；混合交通网络中的出行分布与交通分配模型；交通网络中的动态行为模型；元胞自动机模型、格气模型等在交通行为模拟中的应用；信息技术对网络中的巨量行为的影响模型；公共场所逃生行为的动力学模型。

(2)经济金融行为建模与分析。中国IGEM模型，分析能源、环境、贸易、税收等对经济增长的影响；基于极值理论的货币危机识别模型与方法；多变量极值模型研究；金融市场临界行为建模和突变的预警研究；金融市场的非线性复杂

行为研究；基于行为的具有隐含期权的保险产品定价研究、公司投资行为研究、公司并购行为研究、公司融资行为研究、公司资本结构研究等。

(3)生产管理行为建模与分析。On-line 计划与重调度；生产计划执行中的应急管理与干扰管理；基于需求信息管理的供应链稳定性研究；基于参与和体验的制造和服务混合系统管理研究。

(4)统计与系统管理理论和技术。变量筛选、模型分类与建模自动化研究；符号数据的多元分析代数理论体系；成分数据多元分析的整体建模技术；函数数据的多元分析建模方法；Hilbert 空间的多元分析方法；有关应用案例。

本创新群体的研究成果主要以论文形式发表，近年来在 *Transportation Research*(*Part B*,*Part E*)、*Operations Research Letters*、*Journal of Advanced Transportation*、*European Journal of Operational Research*、*Physical Review E*、*Physica A*、*Network and Spatial Economics*、*Review of Income and Wealth*、*Production Planning and Control*、*Inter J Production Economics*、*Computer & Industrial Engineering*、*Computer & Operations Research*、*Supply Chain Management*:*An International Journal*、*Quantitative Finance*、*International Journal of Finance & Economics*，以及《经济研究》、《经济学季刊》、《管理科学学报》等刊物上发表了大量学术论文。论文毕竟受篇幅限制，难以详细描述研究过程和全部结果，而书是最系统的发表方式之一，在科学出版社的大力支持下，我们将创新群体的研究成果以“复杂社会经济行为建模与管理研究丛书”的形式出版，让更多的读者受益。

本丛书的作者是本创新群体成员及其合作者，每本书重点研究一个问题，要求体现新、深、精，这也是本创新群体的一贯作风，即求真、务实、扎实。我们愿意将自己的成果奉献给大家，在共同分享研究心得的同时，更希望得到广大读者的宝贵意见，为繁荣学术、服务社会作出新的贡献。

黄海军

2011 年 5 月 3 日

前 言

多元回归、判别分析以及主成分分析都是人们十分熟悉的多元数据分析模型。在信息技术飞速发展的今天，数据分析技术的发展也面临着新的机遇与挑战。在许多应用研究中，系统中的数据规模变得越来越大。与此同时，参与系统分析的模型数量也急剧膨胀。面对这样的问题，本书重点讨论了变量筛选、模型分类、模型预测以及自动化建模等技术方法，以便更高速、高效、高质地解决大规模的数据系统的建模和预测分析等问题。

这里所说的变量筛选，主要是指在建模过程中，对变量集合中的信息进行更有效的筛选或综合，尽量删除模型中的冗余变量和无用信息，从而降低所分析问题的复杂程度。为了说明变量筛选的重要意义，本书首先分析了变量多重相关对多元回归、主成分分析以及偏最小二乘回归等几种有代表性的多元分析模型的影响方式。然后，基于 Gram-Schmidt 变换，分别研究了无导师模型与有导师模型的变量筛选方法。

模型分类的主要目的是，在模型数量巨大的情形下，将具有一定相似度的模型划分成一类，然后再对各类模型分别进行研究和建模，这样可以由海量模型化简到有限个模型类，从而极大地降低建模的工作量。为此，本书特别定义了多元分析建模的本征信息，并据此提出了多元分析模型的分类方法。而作为拓展性研究，书中还给出了多元回归与判别分析的模型预测方法。这些模型预测方法可以在不需要未来样本数据的情况下，预测未来时刻模型中的参数和评价模型精度的主要指标。

此外，本书还结合本征信息的概念以及变量筛选技术，讨论了多元回归的自动化建模方法。该方法可以在尽量不依靠人工干预的前提下，识别模型的非线性形式，从而显著提高建模的速度和效率。在本书最后一章，我们还以投入产出表的预测建模问题为例，来说明本书方法论研究的意义与作用。

在相关研究工作中，本书获得国家自然科学基金项目(编号：70771004)、国家自然科学基金重点项目(编号：71031001)、国家自然科学基金创新研究群体科学基金项目(编号：70521001)、国家自然科学青年基金项目(编号：71001110)的支持。作者在该领域的研究还得到法国巴黎高等商学院 M. Tenenhaus 教授和国

际统计学会前副主席 G. Saporta 教授的支持与帮助。1994 年，在 Tenenhaus 教授的引导下，我们开始对偏最小二乘回归方法进行研究，由此开始长期关注系统信息的综合与筛选问题。2007 年，作者应 Saporta 教授邀请在法国国立科学技术与管理大学工作，其间，进一步探讨了变量多重相关性对一些现代模型的影响方式，并且研究了基于 Gram-Schmidt 变换的变量筛选方法。2008 年，Saporta 教授在北京航空航天大学工作期间，在他的指导下，我们又完成了自动化建模的技术框架。所以，在本书出版之际，向两位教授表示最诚挚的谢意！

我们还要感谢国家创新研究群体的项目负责人黄海军教授以及其团队所有成员！他们在这六年多的团队合作与交流中，给予了我们极为重要的指导和帮助。感谢任若恩教授长期以来在学术上的无私帮助和工作上的鼎力支持！感谢国家自然科学基金重点项目(编号：71031001)团队的全体老师和同学对本书理论与应用研究方面的重要贡献！最后，衷心感谢北京航空航天大学经济管理学院的老院长顾昌耀教授和我的导师冯允成教授引导我们走上科学研究的道路！

本书的研究内容，很多是直接来源于实验室同学们的论文研究工作。除本书作者的研究外，还包括了龙文、叶明、仪彬、李岩、王兰会等同学博士论文中的部分内容，以及王劼、陈梅玲、王成、卢珊、夏棒、赵传斌等同学的相关研究成果。正是由于他们的勤奋努力，本书的内容体系才能最终形成。此外，还有很多实验室同学(张瑛、冯甲策、寇薇等)的研究工作，也与本书的内容有一定的关联，感兴趣的读者可以通过相关文献来更全面地了解他们的研究过程和研究内容。需要特别指出的是，陈梅玲同学在最初整理本书的文稿素材时投入了大量的时间和精力，赵传斌同学也对本书的排版等工作给予了重要的帮助。感谢他们的付出！

由于作者的水平有限，书中难免存在缺点和不足，敬请读者批评指正。

作　者

2013 年 5 月

目 录

第1章

绪　论

1.1　引　言

在信息技术飞速发展的今天，统计数据分析理论与方法的发展正面临着许多新的挑战。例如，在许多企业、银行、政府机构的应用项目中，人们经常需要在很短的时间内快速建立成百上千甚至是成千上万个统计模型，或者需要对大量的统计模型进行解释和分析。例如，对众多地区逐年 GDP(groups domestic product，即国内生产总值)的增长模式进行辨识，对工业企业的各种产品的销售进行预测等。这时，“时间”就会成为人们评价建模分析工作的重要因素。在此类应用需求的激励与推动下，如何合理地设计一个标准化的建模过程，以及构建自动化的模型拟合器，从而实现快速建模的目的，这一研究问题引起了人们的极大兴趣。反映到数据分析领域，人们已经开始把研究视角从海量“数据挖掘”拓展到海量“模型挖掘”。由此出现了工业化分析(industrialization of analysis)的概念(Fogelman-Soulié，2008)，这为现代统计数据分析开辟了一个新兴的研究领域。

工业化分析(或称工业化建模)的目的是要在一个较短的时间内，快速建立一大群数量可观的模型。其实，很早以前，这样的业务需求就已经见诸许多企业的管理实践。例如，在 20 世纪 70 年代，美洲航空公司发展的分配与计划系统，就可以提供从咖啡机到起落架的超过 5000 种备件的需求预测。这些备件的保障供应对一架飞机的正常运行都是至关重要的。从库存管理的角度来看，备件短缺会导致航班取消或延误，带来很高的成本；但过于保守地存储过多的备件，又会导致企业积压资金并提高库存成本。由此可见，准确预测这 5000 多种备件的需求

量将会为企业节约大量的资金，并有效地保证飞机运行。为了解决这一问题，美洲航空公司曾采用回归模型，建立月度备件更新数量关于月度飞行小时的函数关系。试想一下，如果大多数回归模型都是非线性的，并且它们的非线性形式都是未知的，那么采用以往手工建模的方式，对每个模型的非线性形式进行反复的尝试性分析，是很难在短时间完成的。从类似的应用需求可以看出，工业化建模技术的发展，必然会极大地推动经济管理理论与实践的发展。

在建模过程中提到的工业化概念，可以简单地概括为合理化、批量化、自动化以及适度的质量控制。这里首先解释“合理化”的概念。追溯到19世纪欧洲工业化初期，泰勒(Taylor，1911)创建的科学管理原理曾为欧美工业化的飞跃式发展提供了重要的理论支撑。当时，为了确定合理的工人作业时间，他通过对熟练工人的工作过程的观察，去除那些冗余的动作，归纳出一套最优的工作方法。然后测定每个动作所需要的时间，制定出合理的工时定额。泰勒这套工作思路对我们的启示是：提高建模效率，首先应该尽可能摒除那些不必要的工作内容。大家知道，多元统计分析处理的是多变量问题，如果分析的变量过多，就会增加所分析问题的复杂性，降低分析的精度。而在实际工作中，人们所选用的诸多变量之间往往存在着较强的相关性，这其中存在着大量的冗余信息。从这个意义上说，进行高效率的数据简约处理，对变量集合中的信息进行筛选或综合，就成为多元分析工业化建模中的关键技术。

批量化也是工业化建模过程中必不可少的重要环节。亚当·斯密《国富论》的第一篇一开头就谈到了“论分工”，他详细讨论了专业化分工对提高生产效率的巨大作用。这说明对众多任务分门别类后再进行处理，要远比处理大量繁杂的个体任务快捷高效。因此，当面对成千上万个建模任务时，如果我们能够给出对这些模型群进行分类的方法，并识别不同类型模型的主要特征，就可以批量化地进行建模，显著提高建模速度。例如，在对中国的城市经济发展水平进行预测时，要分别建立600多个模型。然而如果通过分类，将经济发展特征以及经济发展趋势一致的城市分成一类，然后再对每一类中的“代表城市”进行建模预测，则通常可以极大地降低建模的工作量。当然，由于每类模型只需要建立一个代表性的模型，因此必须确保该模型对类内所有其他模型的代表性，这样才能控制模型群的整体质量。综上所述，对多元分析模型群的分类技术将为工业化建模的发展提供重要的理论和技术支撑。

工业化建模的第三个重要的概念是“自动化”。没有自动化的过程，就无法真正脱离人工手动的建模方式，也就无法实现高速、高效的建模任务。在本书所述的主要工作环节中，都将体现自动化的意识。例如，在变量筛选过程中，我们可以根据事先指定的精度要求，自动确定计算的停止准则，并保证变量选择方案的最优性质；在模型分类过程中，由于采用了一种改进的Squeezer算法来取代K-

means 算法，因此，我们可以在控制每类模型的误差范围的同时，全自动化地进行模型分类。在本书中，对自动化概念体现最为突出的地方是对非线性回归模型形式的自动识别技术。本书将拟线性回归技术与变量筛选技术相结合，实现了对任意多元非线性回归模型形式进行有效辨识的目标。

本书将变量筛选、模型分类以及自动化建模技术结合在一起，就形成了一个工业化建模的程序框架。在第 8 章，我们将讨论投入产出表的预测建模技术。在这个模型系统中，曲线分类与自动化建模是重要的组成部分。我们将通过这个实用问题，来说明本书方法论研究的意义与作用。

1.2 本书的内容结构

本书以工业化建模任务为主线，讨论了回归模型的变量筛选、模型分类和自动化建模问题。以此为核心，还适当扩充了其他相关的研究内容。

书中的第 2～5 章，都是讨论数据降维和变量筛选的章节。所谓数据降维，是指要寻找一个可以反映事物的本质规律的、低维度的变量集合，用它来表示原始的高维变量集合。Carreira-Perpinán(1997)曾对数据降维问题做出概要性综述，他指出，之所以能对高维数据进行降维，是因为在原始数据集合中常常包含大量的冗余信息，通常剔除一部分冗余信息，有可能找到一组对原数据更经济的表达方式。数据降维的概念对于每一个熟悉多元分析应用的人来说都并不陌生。例如人们常用的主成分分析、Fisher 判别模型、偏最小二乘回归以及经典的逐步回归等，它们在方法论上有一个统一的特征，即首先把原始的高维空间降到低维，然后再做进一步的分析。

为了说明各种数据降维方法的区别，Cunningham(2008)曾对各种降维技术做了大致的分类，如图 1.1 所示。他将多元分析模型分为有导师(supervised)模型和无导师(unsupervised)模型。由于工作任务不同，这两类模型在降维过程中所秉持的原则是不一样的。例如，有导师模型是有因变量的模型(如回归模型、判别分析模型等)。针对这类模型的降维问题，其理想目标是删除所有对因变量没有解释作用的变量，挑选出对因变量解释能力最强的自变量，而且如果在自变量集合中存在较强的相关性，最好能删除所有的冗余信息。另一类多元分析模型被称为无导师模型，这类模型的主要特征是在原始数据中没有自变量和因变量之分(如主成分分析、典型相关分析等)。在降维过程中，这类模型要遵守的优化原则是降维后的简约变量集合能够最好地反映原始变量集合中的总信息量。

	有导师	无导师
特征提取	主成分回归 偏最小二乘回归	主成分分析 聚类分析
变量筛选	回归中的变量筛选 Fisher 判别分析	主基底分析

图 1.1 数据降维方法的分类(与本书相关的几种)

本书的第 4 章重点研究了一种被称为主基底分析的模型，给出了无导师模型的变量筛选方法。而第 5 章，则以回归分析和判别分析这两个有导师模型为例，讨论了一类有导师模型的变量筛选问题。

对降维方法进行分类的另外一个办法，是根据它们的工作方式，将其分成特征提取(feature transformation)类方法和变量筛选(variable selection)类方法。最常见的特征提取方法有主成分分析、Fisher 判别分析、典型相关分析、偏最小二乘回归等。这些方法的共同特点是依照某种最优化原则，将 p 个原始变量 $\boldsymbol{x}_1$，$\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p$ 综合为 m 个新变量(通常被称为“成分”) $\boldsymbol{F}_1,\boldsymbol{F}_2,\cdots,\boldsymbol{F}_m(m<p)$，然后再利用这些成分进行后续的分析工作。通常情况下，m 都会远远小于 p，所以实现了将高维空间降到低维的目的。然而需要强调的是，因为每一个成分都是原始变量的线性组合，所以这类方法并不具备变量筛选的功能。

变量筛选方法的工作思路不同于特征提取方法。它是在原始的变量集合中真正筛选出一个子集，以此来代表原变量集合。变量筛选在有导师的模型中已经有很多应用的先例，如普通最小二乘回归中的变量筛选方法，就是把自变量对因变量的解释能力作为变量筛选的目标，来设计搜索算法和停止准则。在本书的第 3～5 章，我们将重点讨论变量筛选的理论方法以及应用问题。

本书的第 2 章是一个特殊的章节，它所探讨的问题是特征提取方法是否能有效地解决变量集合中的多重共线性问题。对于许多应用研究人员来说，一提到数据降维，总会立刻联想到主成分分析。鉴于主成分分析能够在信息损失最小的前提下，将高维空间降到低维，并且由于主成分之间是线性无关的，一些研究人员在使用主成分分析方法时就显得不够谨慎。他们会比较随意地选择一大群变量，然后期待通过主成分分析，把高维空间降到低维的同时，自动消除变量之间的多重相关性。这种错误观念后来又延伸到对偏最小二乘回归等方法的应用工作中。由于盲目相信偏最小二乘回归方法在消除变量多重共线性中的作用，有些工作人

员甚至会刻意使用上百个变量来建立一个预测模型。事实上，由于许多特征提取的降维方法在计算过程中使用了较复杂的数学变换，而且经常可以把一个多重相关的变量集合变换成线性无关的新变量集合，所以一些分析人员就会认为，这些降维方法可以完全消除变量之间的多重共线性，于是在实际应用时，便宁滥勿缺地选择一个很大的变量集合，然后等待这些降维方法去自动地发挥作用。

不过，常识性的问题是根据物理学中的能量守恒定律，当某种数学变换把一个严重多重共线的变量系统变成相互独立的变量系统后，原始的冗余信息到哪里去了呢？它是否依然存在？是否会换一副面孔、以另一种形式来影响最终的分析结论呢？为了回答这些问题，本书在第 2 章中详细讨论了变量多重共线性对特征提取类建模方法的影响。研究表明，主成分分析、普通最小二乘回归以及偏最小二乘回归等方法无论在信息筛选还是在参数估计方面，都会以不同形式、在不同程度上受到变量多重相关的影响。也就是说，当研究人员有意识或无意识地增加某个变量集合中的多重相关性时，便会人为增强某些变量在这些模型中的地位与作用，这使得模型的合理性受到影响，而对模型含义的解释也将是不准确和非客观的。

数据降维的另一种思路是进行变量筛选。在经典统计学中，已经存在诸多的变量筛选方法。本书的第 3 章，将重点讨论一类新的变量筛选方法，即基于 Gram-Schmidt 变换的变量筛选方法。Gram-Schmidt 变换是由 Gram 和 Schmidt 提出来的，其特点是能够将一组线性无关的变量构造成直交变量。王惠文等(2008a)在讨论多重相关条件下的变量筛选和建模方法时指出，可以利用 Gram-Schmidt 正交化过程，把自变量集合变换成若干直交变量与若干零变量的组合，并基于这个思想提出了 Gram-Schmidt 回归。这是一种将 Gram-Schmidt 过程运用在有导师变量筛选的方法。之后，王惠文等在 2012 年又给出一种快速 Gram-Schmidt 回归方法，可以成批量地删除冗余变量，进一步提高了变量筛选的速度。随后，我们又将该方法推广到非线性回归、判别分析、主成分分析等有导师模型以及无导师模型的变量筛选等研究当中，使得所建模型在有效性和可解释性等方面都得到了较大的改善。

本书第 6 章的核心目的是为研究多元分析模型的分类方法奠定理论基础。笔者在该章中指出，尽管多元分析方法是形形色色的，但是它们却有一个共同形式的本征信息，这就是原始数据的协方差矩阵。这里提到的本征信息是指对于两个多元分析模型，只要它们的本征信息是一致的，那么这两个模型的系数以及各种模型评价参数的计算结果就是完全相同的。从这个结论出发，不难发现，对多元分析模型群进行分类的关键技术就是，如何对协方差矩阵进行聚类分析。

在给出上述理论结论后，第 6 章还拓展性地探讨了多元分析模型的预测问题。随着长时间数据信息的收集与积累，对于截面数据表的持续观察形成了纵向

数据表(longitudinal data)。由于增加了时间维度，其数据结构与截面数据表相比发生了质的变化，而研究内容也更加丰富。从 20 世纪 60 年代以来，研究人员就开始对纵向数据表进行分析研究并取得了很多著名的研究成果，这些方法为对纵向数据表中的知识挖掘和建模研究提供了许多有价值的研究工具。在对纵向数据表的研究中，有一个极富挑战性的课题，即如何利用此类数据对未来多元模型的参数和性质进行预测估计。也就是说如何根据一系列按时间顺序收集的数据表，建立多元分析的预测模型，从而在不需要未来样本数据的情况下，预测未来时刻多元分析模型中的参数，并推测评估模型精度的主要指标。在该章中，将以多元线性回归和 Fisher 判别分析为例，研究多元分析的模型预测理论与方法的问题。

本书的第 7 章主要针对回归分析模型，从建模的批量化和自动化两个方面，提出相关的自动聚类与自动建模的方法。书中首先介绍了大规模曲线的自动聚类方法，这相当于是一元非线性回归模型的聚类问题。然后，基于多元回归模型的本征信息，讨论了多元回归模型的聚类问题。为了实现自动化建模的目的，我们采用拟线性回归的方法将模型从非线性空间映射到线性空间，然后再与变量筛选方法相结合，给出对多元非线性模型形式进行自动辨识的方法。最后在第 8 章，作为一个应用实例，介绍了投入产出表的预测建模方法。

综上所述，下面再分章节陈述一下本书的内容。

第 2 章，介绍变量多重相关性对特征提取类建模方法的影响。本章首先简要介绍了多重相关性的概念和产生原因，然后按照无导师和有导师两类模型，分别讨论变量多重相关性对模型的影响方式。在无导师模型方面，主要介绍了多重相关性对于主成分分析的影响，指出主成分分析并不能摒除数据系统中的冗余信息。在有导师模型方面，我们将主要论述多重相关性对普通最小二乘回归、主成分回归以及偏最小二乘回归分析的影响。

第 3 章，介绍 Gram-Schmidt 变换及其反变换的方法，说明经 Gram-Schmidt 变换所删除信息的含义。然后，给出对系统中的冗余变量进行假设检验的方法。这些内容也为后续章节研究变量筛选方法奠定了理论基础。

第 4 章，给出基于 Gram-Schmidt 过程的无导师变量筛选方法。

第 5 章，主要介绍基于 Gram-Schmidt 过程的有导师变量筛选方法，包括基于 Gram-Schmidt 过程的多元回归分析和判别分析。

第 6 章，重点介绍了多元分析模型的本征信息概念，为后续模型分类做理论基础准备。作为拓展性内容，本章还讨论了多元线性回归、Fisher 判别分析等多元分析模型的预测方法。

第 7 章，讨论了自动化回归建模的方法和过程。书中首先研究了大规模曲线自动聚类方法；其次给出了多元线性回归模型的自动聚类方法。在此基础上，讨论了在回归建模过程中的自动化模型辨识方法。

第 8 章，以投入产出表的预测建模问题为例，说明曲线分类与自动化建模的意义和作用。

在介绍上述方法的同时，本书在每章的方法研究后面都给出仿真研究或者案例分析，帮助读者更好地理解这些方法的计算过程，以便在实际工作中操作和运用。

1.3　数据表的基本知识

在本书中，关于各种模型方法的讨论，所处理的对象都是通过对多元变量总体进行抽样观测而得到的数据表，建模过程都是在对数据表进行变换和计算的基础上完成的。因此，作为最基本的准备知识，本节拟简单介绍数据表的构成、基本概念和一些技术处理。

在多元统计分析中，研究的是样本点×定量变量类型的平面数据表。假设有 p 个变量 $x_1,x_2,\cdots,x_p$，对它们分别进行 n 次观测，则由此所构成的数据表 $\boldsymbol{X}$ 可以写成一个 $n\times p$ 维的矩阵

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=\begin{bmatrix} \boldsymbol{e}_1^{\mathrm{T}} \\ \boldsymbol{e}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{e}_n^{\mathrm{T}} \end{bmatrix}=[\boldsymbol{x}_1 \quad \boldsymbol{x}_2 \quad \cdots \quad \boldsymbol{x}_p]$$

其中，$\boldsymbol{e}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}\in\mathbf{R}^p(i=1,2,\cdots,n)$，$\boldsymbol{e}_i$ 被称为第 i 个样本点。在数据表 $\boldsymbol{X}$ 中有 n 个样本点，样本点名为 $\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n$，每个样本点均用 p 个指标变量来描述。

在数据表 $\boldsymbol{X}$ 中，$\boldsymbol{x}_j=(x_{1j},x_{2j},\cdots,x_{nj})^{\mathrm{T}}\in\mathbf{R}^n(j=1,2,\cdots,p)$，$\boldsymbol{x}_j$ 被称为第 j 个变量，表示所有样本点在第 j 个指标上的取值分布，于是数据表 $\boldsymbol{X}$(有时也称为矩阵 $\boldsymbol{X}$)包含了全部变量的所有观测值。

1.3.1　样本点空间

从数据表 $\boldsymbol{X}$ 可以看出，样本点 $\boldsymbol{e}_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}\in\mathbf{R}^p(i=1,2,\cdots,n)$。$\mathbf{R}^p$ 是一个 p 维欧式空间，在 $\mathbf{R}^p$ 中可以定义内积

$$\langle \boldsymbol{e}_i,\boldsymbol{e}_k\rangle=\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{e}_k=\sum_{j=1}^{p}x_{ij}x_{kj} \tag{1.1}$$

由内积可以定义向量 $\boldsymbol{e}_i$ 的模长 $\|\boldsymbol{e}_i\|$，有