HZ BOOKS
华章教育

普通高等院校计算机课程规划教材

MATLAB Data Analysis Methods

MATLAB 数据分析方法

李柏年 吴礼斌 主编　张孔生 丁华 参编

机械工业出版社
China Machine Press

本书以专业理论为指导，以应用软件为工具，以建立数学模型为方法，以解决实际问题为目的，以提高学生的创新能力为宗旨，主要介绍了样本数据的处理方法、线性回归模型与非线性曲线拟合、主成分分析与典型相关分析、判别分析方法、聚类分析以及数值模拟的方法。

本书具有以下特点：

- 既注重数据分析的原理介绍，又注重MATLAB程序的编写。大部分例题给出面向过程的MATLAB程序，有利于学生学习数据分析的原理与提高使用软件的能力。
- 理论与实践相结合，每一章设计了综合性的实验内容，实验理论密切联系社会实际，有利于培养学生分析问题与解决问题的能力，增强学生的社会责任意识。
- 为方便读者，本教材提供课件和例题程序源代码，有需要的读者可以到华章网站（http://www.hzbook.com）下载。

客服热线：(010) 88378991, 88361066
购书热线：(010) 68326294, 88379649, 68995259
投稿热线：(010) 88379604
读者信箱：hzjsj@hzbook.com

华章网站 http://www.hzbook.com

网上购书：www.china-pub.com

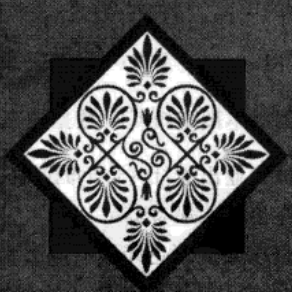
封面设计：杨宇梅

普通高等院校计算机课程规划教材

MATLAB Data Analysis Methods

MATLAB 数据分析方法

李柏年 吴礼斌 主编 张孔生 丁华 参编



机械工业出版社
China Machine Press

数据分析是用适当的统计方法对各种数据加以详细研究和概括总结的过程，已成为当代自然科学和社会科学各个学科研究者必备的知识。MATLAB是一套高性能的数值计算和可视化软件，是实现数据分析与处理的有效工具。本书介绍数据分析的基本内容与方法，应用MATLAB软件既面向对象又面向过程地编写实际数据分析程序。全书共分7章，主要内容包括：MATLAB基础、数据描述性分析、回归分析、判别分析、主成分分析与典型相关分析、聚类分析、数值模拟分析。

每章末精心编写习题供读者练习，此外每章还安排了紧密联系实际的综合性、分析性实验内容。

本书适用于计算机科学与技术、信息与计算科学、统计学等专业的本科生，还可作为相关专业本科生选修课程教材，并可供硕士研究生以及科技工作者参考。

图书在版编目（CIP）数据

MATLAB数据分析方法 / 李柏年，吴礼斌主编. —北京：机械工业出版社，2012.1
（普通高等院校计算机课程规划教材）

ISBN 978-7-111-36287-6

Ⅰ. M…　Ⅱ. ①李… ②吴…　Ⅲ. Matlab软件—高等学校—教材　Ⅳ. TP317

中国版本图书馆CIP数据核字（2011）第223500号

机械工业出版社（北京市西城区百万庄大街22号　邮政编码　100037）
责任编辑：朱秀英
北京市荣盛彩色印刷有限公司印刷
2012年1月第1版第1次印刷
185mm×260mm · 12.5印张
标准书号：ISBN 978-7-111-36287-6
定价：29.00元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88378991；88361066
购书热线：（010）68326294；88379649；68995259
投稿热线：（010）88379604
读者信箱：hzjsj@hzbook.com

前 言

随着信息技术的飞速发展，数据的产生和存储达到了空前繁荣的阶段。从数据中提取有用信息，是数据分析的基本目的之一。结合应用软件介绍基本的数据分析方法也是对传统教学模式的改革。我们以专业理论为指导，以应用软件为工具，以建立数学模型为方法，以解决实际问题为目的，以提高学生的创新能力为宗旨，在积累多年教学实践经验的基础上编写了本书。

本书共7章，其中，第1章对MATLAB软件功能及使用方法进行了介绍，包括软件的基本界面、数据矩阵及运算、程序的编写与文件操作等。第2章介绍了数据描述性分析，包括基本统计量与数据可视化、数据分布及检验、数据变换。第3章介绍了回归分析，包括一元回归模型、多元线性回归模型与逐步回归。第4章介绍了判别分析，包括距离判别分析、Bayes判别分析等。第5章介绍了主成分分析与典型相关分析，注重主成分分析实际应用典型案例的介绍。第6章介绍了聚类分析，包括距离聚类、谱系聚类与K均值聚类、模糊均值聚类等。第7章介绍了数值模拟分析，包括蒙特卡罗方法与应用、BP神经网络及应用。

本书具有以下特点：

(1) 注重数据分析的原理介绍，更注重MATLAB程序的编写。大部分例题给出面向过程的MATLAB程序，这有利于学生学习数据分析的原理与提高使用软件的能力。

(2) 理论与实践相结合。每一章设计了综合性的实验内容，实验密切联系社会实际，这有利于培养学生分析问题与解决问题的能力，增强学生的社会责任意识。

使用本书的读者应有一定的计算机高级语言编程基础，学习过高等数学、线性代数、概率统计等课程。我们将全书的例题程序汇编放在华章网站（http：//www. hzbook. com），以方便读者学习。

本书由李柏年、吴礼斌主编并统稿，张孔生、丁华参编。具体分工如下：李柏年编写第2章、第5章及各章实验；吴礼斌编写第1章、第3章和第7章；张孔生、丁华编写第4章和第6章。

在本书编写过程中，南京大学徐洁磐教授始终给予了热情的关心与支持，提出过很多宝贵的建议，在此向他表示衷心的感谢！

限于我们的水平，书中不当之处在所难免，敬请读者批评指正。

编 者

2011年7月

教学建议

在教学过程中，一要重视数据分析原理的介绍，二要重视 MATLAB 程序编写的算法分析，三要重视每章的综合性实验教学。学生应具有计算机高级语言编程基础，学习过高等数学、线性代数、概率统计等课程。建议总教学时数为 54 学时，其中综合实验为 24 学时。建议课堂教学在计算机多媒体机房内完成，实现“讲与练”结合，实验课要求学生提交实验报告。具体各章的教学时数、内容和要求可作如下安排：

第 1 章　MATLAB 基础（6 学时，其中 2 学时实验）

教学内容：

MATLAB 与数据分析；MATLAB 的基本界面操作；矩阵的基本运算；MATLAB 编程与 M 文件。

教学要求：

熟练掌握 MATLAB 的基本界面操作；理解软件的运算符、操作符、基本数学函数命令等的功能与调用格式；掌握矩阵的运算；熟练掌握选择、循环语句的编程；掌握建立 M 文件的方法。

第 2 章　数据描述性分析（8 学时，其中 2 学时实验）

教学内容：

基本统计量（如均值、方差、分位数等）与数据可视化；数据分布与检验（一元与多元数据）；数据变换（无量纲化、box-cox 变换等）。

教学要求：

熟练掌握利用 MATLAB 软件计算基本统计量与数据可视化；掌握 JB 检验与 Lilliefors 检验关于数据的正态性检验；掌握协方差矩阵相等的检验方法；理解数据变换。

第 3 章　回归分析（8 学时，其中 4 学时实验）

教学内容：

一元回归模型（线性与非线性回归模型）；多元线性回归模型；逐步回归分析。

教学要求：

理解回归分析的原理；熟练掌握 MATLAB 回归分析的命令；掌握非线性回归的基本方法以及 MATLAB 实现；掌握逐步回归的 MATLAB 方法。

第 4 章　判别分析（8 学时，其中 4 学时实验）

教学内容：

距离判别分析；贝叶斯判别分析；判别准则的评价。

教学要求：

理解判别分析的原理；熟练掌握 MATLAB 软件进行距离判别与贝叶斯判别的方法与步

骤；掌握判别分析的回代误判率与交叉误判率的计算；掌握解决实际判别问题的建模方法。

第5章　主成分分析与典型相关分析（8学时，其中4学时实验）

教学内容：

主成分分析的原理（总体主成分的定义、计算、性质，样本主成分的计算方法）；主成分分析的应用（基于主成分分析的综合评价、分类、信号分离等）；典型相关分析（原理、典型相关系数计算、检验，样本数据典型相关变量）；典型相关分析应用实例。

教学要求：

理解主成分与典型相关分析的原理；熟练掌握利用MATLAB进行主成分分析的计算步骤；掌握MATLAB进行典型相关分析的计算步骤；掌握具体实际问题典型相关分析结果的合理解释。

第6章　聚类分析（8学时，其中4学时实验）

教学内容：

距离聚类分析（向量距离、类间距离）；谱系聚类与K均值聚类；模糊均值聚类（模糊C均值聚类，模糊减法聚类）；聚类的有效性。

教学要求：

理解聚类的思想与原理；熟练掌握MATLAB关于各种样品距离与类间距离的计算方法；会作谱系聚类图；掌握MATLAB各种聚类命令的使用方法；掌握聚类效果分析方法及程序的实现。

第7章　数值模拟分析（8学时，其中4学时实验）

教学内容：

蒙特卡罗方法与应用（基本思想，MTLAB的伪随机数，应用实例）；BP神经网络与应用（神经网络的概念，BP神经网络，MATLAB神经网络工具箱，BP神经网络的预测与判别）。

教学要求：

理解蒙特卡罗方法；掌握利用MATLAB生成伪随机数的方法；掌握伪随机数的应用；理解神经网络的基本思想；掌握利用MATLAB实现神经网络的预测与判别。

目　录

第1章 MATLAB基础

本章主要介绍MATLAB软件的一些入门知识，包括MATLAB界面及其基本操作、变量与函数、运算符与操作符、数据矩阵的输入与输出、符号运算、M文件与编程等，为读者学习以后各章打下基础。

1.1 数据分析与MATLAB

1.1.1 数据分析概述

1. 数据分析的概念

数据分析是指用适当的统计方法对收集来的大量第一手资料和第二手资料进行详细研究，提取有用信息并形成结论，以求最大化地开发数据资料的功能与发挥数据的作用。在统计学领域，有些人将数据分析划分为描述性统计分析、探索性数据分析以及验证性数据分析，其中，探索性数据分析侧重于在数据之中发现新的特征，而验证性数据分析则侧重于已有假设的证实或证伪。

2. 数据的来源与分类

数据是数据分析的关键之一。数据也称观测值，是实验、测量、观察、调查等的结果，常以数量的形式给出。数据按照不同的标准进行分类，可分为：观测数据与试验数据、一手数据与二手数据、时间序列数据与横截面数据等。

1）观测数据与试验数据。观测数据（observational data）是在自然的未被控制或者不可控制的条件下观测到的数据，如社会商品零售额、消费价格指数、汽车销售额、降雨量等。利用这类数据进行的观测研究，是观测所研究的个体，并度量感兴趣的变量，但并不会影响其回应。试验数据（experimental data）是在人工干预和操纵的条件下产生的数据。这种数据通常来自于科学与技术实验。例如，了解不同的药物成分对某种疾病的治疗效果有什么不同，试验在药物成分不同的条件下产生相应的治疗效果数据。这些药物成分数据和治疗效果数据就是试验数据。将数据分为观测数据和试验数据是基于所观测的对象是在自然的还是可控的实验条件下产生的。观测研究是被动的数据收集方式，我们只观察、记录或度量，但是不干扰。许多自然现象和社会现象无法控制，只能通过观测获得数据。而实验能够主动产生数据，做实验的人会主动介入，可以把某项处理施加到受试对象，来观察受试对象有何反应，如上述的治疗效果数据。

2）一手数据与二手数据。二手数据是由各种媒体、机构等发布的数据，如证券市场行情、物价指数、耐用消费品销售量、利率、国内生产总值、进出口贸易数据等。对于数据分析人员来说，可以根据研究的问题，从这些数据中加以选择。二手数据是间接得到的。在数

据分析中，有许多数据不能像获取二手数据那样间接得到，而必须进行专门收集、调查或试验才能获得。这种针对特定的研究问题，通过专门收集、调查或试验获得的数据称为一手数据。例如，为制订一家百货商店的营销方案，在这家商店所在的城市抽取近 300 户家庭作为样本进行调查，收集下列数据：对本商店及其竞争对手商店的熟悉程度；家庭成员在各个商店购物的频率；选择百货商店时考虑的因素，如商品质量、种类、退赔政策、服务、价格、店址、商店布局、信用与收款政策；每个商店的偏好评分；被调查者的年龄、性别、受教育程度等。这些数据都是为解决该商店营销问题专门调查收集的，因此是一手数据。一手数据与二手数据是对数据分析人员获取数据的直接与间接两种方式的划分结果。

3）时间序列数据与横截面数据。时间序列数据是对同一研究对象按时间顺序收集得到的数据，如国内生产总值（GDP）、失业率、社会商品零售额等。这类数据是按照一定的时间间隔每日、每周、每月、每季、每年收集的。例如，由每季 GDP 组成的时间序列数据、由每年 GDP 组成的时间序列数据、商店日销售额时间序列数据、商店周销售额时间序列数据、商店月销售额时间序列数据、商店季销售额时间序列数据、商店年销售额时间序列数据。横截面数据是指在同一时点上不同的研究对象的数据的集合，如 2010 年沪深股市上市公司的中期业绩。由这两类数据衍生出合并数据，合并数据中既有时间序列数据又有横截面数据。例如，收集 2000—2010 年 10 个国家的国内生产总值，从每个国家的角度看，每个国家国内生产总值都组成 2000—2010 年的时间序列数据；从每个时点上看，例如，2010 年 10 个国家的国内生产总值组成横截面数据。可以将合并数据理解为横截面数据按时间顺序排列得到的数据集。在合并数据中有一类特殊的数据，称为 panel 数据（panel data），又称纵向数据，即同一个横截面单位在不同时期的数据。例如在一定时期间隔内对同一地区同样的家庭进行调查，以观察其住房和经济状况是否有变化，这样得到的数据就是纵向数据。时间序列数据与横截面数据是数据沿时间与个体两个维度上的视图。

3. 数据分析的过程

数据分析的目的是利用数据来研究一个领域的具体问题。数据分析的过程包括确定数据分析的目标、研究设计、收集数据、数据整理与分析、解释和分析计算结果。

1）确定数据分析的目标。数据分析的目标是分析和解决特定的领域问题，而这个问题可以用量化分析的方法来解决。

2）研究设计。数据分析的研究设计，是根据数据分析的目标寻求解决方案。一般而言，数据分析是用量化分析的方法对现象进行描述、解释、预测与控制。一个特定的领域问题要先转化为数据分析问题。为此，首先要进行量化研究设计，即确定用什么量化研究方法来进行该问题的研究及怎样研究。常用的量化研究方法有调查法（用调查或观测得到的样本数据推断总体）、相关研究法、实验法、时序分析法等。在这一阶段确定量化研究的方法，确定数据分析目标的解决方案。

3）收集数据。确定了所要解决的问题的研究设计后，根据所要采用的量化研究方法收集数据。例如，若采用调查法，需要确定具体抽样方法以获取数据；若采用实验法，需要进行实验设计，通过实验来获取数据等。这些是为所要解决的问题专门收集的一手数据。除此之外，通常还需要二手数据。

4）数据整理与分析。数据整理与分析，即利用数据分析方法进行计算和分析。数据分析方法以统计分析技术为主。这一步骤主要是整理数据并应用数据分析方法进行计算和分析。需要以各种软件（SPSS、SAS、Excel、S-Plus 等）为工具，或对特殊分析方法通过编制

程序来进行计算。本书以 MATLAB 为工具进行计算。这里特别要注意分析方法与软件包的结合。

5）解释和分析计算结果。使用各种软件包等工具计算后，会得到一系列结果，包括各种图表、数据等。说明、解释和分析这些结果，或利用计算结果检验各种假设、预测、控制等，从而最终解决所研究的问题。这一阶段完成数据分析报告，同样是数据分析全过程非常重要的一环。

上述阐明了数据分析的一般过程。各个步骤之间常常需要互相反馈调整。

1.1.2　MATLAB 在数据分析中的位置和作用

从数据分析的整个过程来看，软件的使用主要是在第四阶段，即数据整理与分析阶段。软件所起的作用主要是整理、计算、绘制图表等。

MATLAB 是一套高性能的数值计算和可视化软件，它集矩阵运算、数值分析、信号处理和图形显示于一体，构成了一个界面友好、使用方便的用户环境，是实现数据分析与处理的有效工具，其中 MATLAB 统计工具箱更为人们提供了一个强有力的统计分析工具。

MATLAB 软件作为数据分析的工具提高了计算能力，使得样本容量扩大，增加了统计推断的正确性，也促进了包含大量计算的多元统计分析等方法的发展和应用。软件的使用还为数据分析过程节约了大量计算时间，提高了数据分析的效率。尽管软件对数据分析起到非常大的作用，但它不能处理数据分析中的其他几个阶段所解决的问题。明确这一点后可以更好地使用软件。确定数据分析的目标，对问题的研究设计，选择统计分析方法，收集数据，解释和分析计算结果，这些都不是软件所能解决的。将一批数据输入计算机，用统计软件包含的许多方法都可以产生结果，但计算机及软件不了解用户要干什么。数据分析的结果是否合理完全取决于用户是否知道自己在干什么以及是否清楚所要解决的问题与解决问题的方法。

因此，本书的宗旨就是将数据分析与 MATLAB 的应用作为一个不可分割的整体来讨论。为了方便不了解 MATLAB 软件使用方法的读者学习，以下对 MATLAB 的基本操作方法作比较系统的介绍。

1.2　MATLAB 简介

MATLAB 是由美国 Mathworks 公司推出的一个科技应用软件，其名字来源于矩阵（matrix）和实验室（laboratory）两词的前 3 个字母。它是一种广泛应用于工程计算及数值分析领域的新型高级语言，可以把科学计算、结果可视化和编程都集中在一个使用非常方便的环境中。自 1984 年该软件推向市场以来，历经 20 多年的发展，现已成为国际公认的最优秀的工程应用开发软件之一。MATLAB 功能强大、简单易学、编程效率高，深受广大科技工作者的欢迎。在国际学术界，MATLAB 已经被确认为是准确、可靠的科学计算标准软件。在许多国际一流学术刊物（尤其是信息科学刊物）上，都可以看到 MATLAB 的应用。

1.2.1　MATLAB 的特点

MATLAB 的特点如下：

1）MATLAB 是一个交互式软件系统，输入一条命令，立即就可以得出该命令的结果。

2）数值计算功能。以矩阵作为基本单位，但无需预先指定维数（动态定维）；按照 IEEE 的数值计算标准进行计算；提供十分丰富的数值计算函数，方便计算，提高效率；命令与数学中的符号、公式非常接近，可读性强，容易掌握。

3）符号运算功能。和著名的 Maple 软件相结合，具有强大的符号计算功能。

4）绘图功能。提供了丰富的绘图命令，能实现一系列可视化操作。

5）编程功能。具有程序结构控制、函数调用、数据结构、输入输出、面向对象等程序语言特征，而且简单易学、编程效率高。

6）丰富的工具箱。工具箱实际上是用 MATLAB 的基本语句编成的各种子程序集，用于解决某一方面的专门问题或实现某一类的新算法。工具箱可分为功能型和领域型。功能型工具箱主要用来扩充 MATLAB 的符号计算功能、图形建模仿真功能、文字处理功能以及与硬件实时交互功能，能用于多种学科。领域型工具箱专业性很强，如控制系统工具箱（Control System Toolbox）、信号处理工具箱（Signal Processing Toolbox）、符号数学工具箱（Symbolic Math Toolbox）、统计工具箱（Statistics Toolbox）、优化工具箱（Optimization Toolbox）、财政金融工具箱（Financial Toolbox）等。

1.2.2 MATLAB 7.0 界面

MATLAB 7.0 是在 MATLAB 6.5 版本基础上升级而来的，和 6.5 版本相比较界面没有太大改变，命令窗口仍然是用户主界面，图形窗口用来显示图形信息和创建图形用户接口（GUI），文本编辑器用来创建和编辑 MATLAB 代码，MATLAB“桌面”菜单用来调整其他一些窗口的位置和可视性，如工作空间管理窗口、帮助窗口、历史命令记录窗口等。

除了以上不变特性以外，MATLAB 7.0 在一些数值表示和操作方法上有了新变化。MATLAB 7.0 添加和修改了一些内核数值算法，能支持各种数据类型的数学运算，而不仅仅是精度类型的数组（这一数据类型曾一度是较早 MATLAB 版本的核心）。更重要的一点是，MATLAB 7.0 的命令解释程序增加了一个加速特性，它将一个循环视为一个整体进行代码解释和代码执行而非逐行处理（MATLAB 6.5 之前的版本就是这样处理的），从而大大提高了循环操作执行的速度。总之 MATLAB 7.0 是 MATLAB 版本演进过程中的一次改进，但几乎所有用 MATLAB 6.5 编写的代码都可以不加修改地在 MATLAB 7.0 中运行。大部分 MATLAB 7.0 新增和改进的特性都是为了使用户在利用 MATLAB 解决问题时取得更高的工作效率而添加的。

1. MATLAB 工作环境

MATLAB 有以下 3 种启动方法：在安装有 MATLAB 7.0 的计算机上，双击 Windows 桌面上的快捷图标；从“开始”菜单的“程序”子菜单中选择“MATLAB”；在 MATLAB 目录中搜索到可执行程序“MATLAB.exe”，双击该程序使之启动。启动后的界面如图 1-1 所示。

图 1-1 大致包括以下几个部分：菜单栏；工具栏；“Command Window”命令窗口；“Workspace”工作空间管理窗口；“Command History”历史命令记录窗口；“Current Directory”当前目录窗口。

主菜单包括“File”、“Edit”、“Debug”、“Desktop”、“Window”和“Help”。

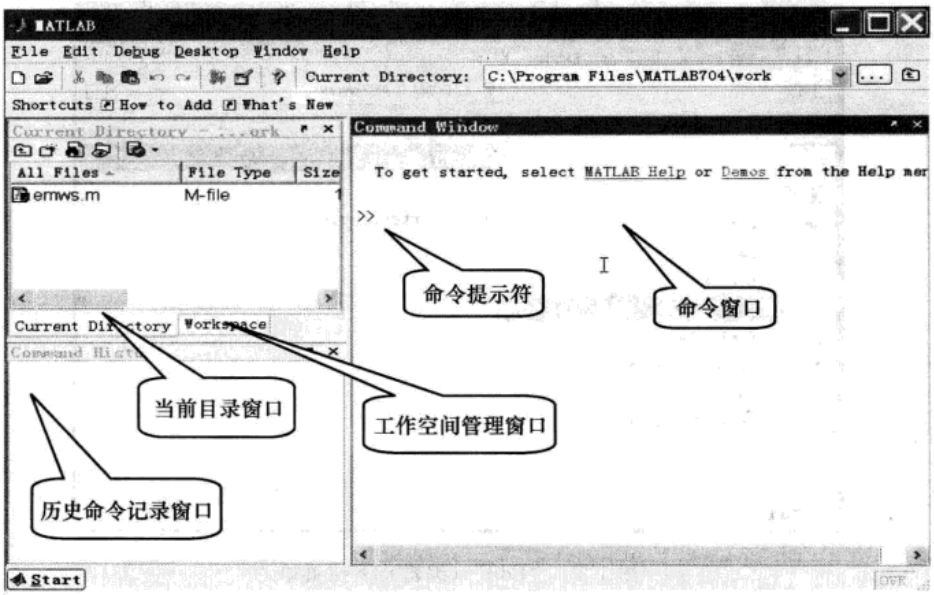


图 1-1　MATLAB 7.0 的默认操作界面

1）“File”（文件）菜单（如图 1-2 所示）。文件菜单除了具有 Windows 一般应用程序所具有的“新建”、“打开”、“关闭”、“退出”、“打印”外，还包括如下菜单项：

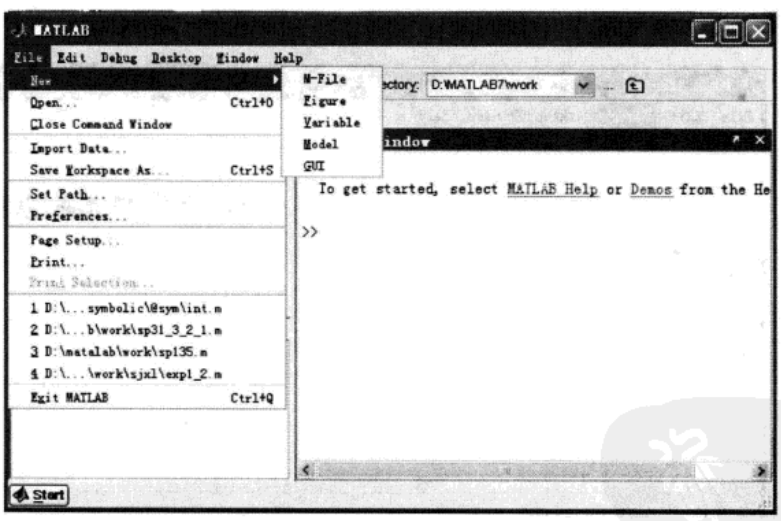


图 1-2　“File”（文件）下拉菜单

“Import Data…”导入有关数据；

“Save Workspace As…”保存工作平台；

“Preferences…”设置部分 MATLAB 工作环境的交互性；

“Set Path…”设置当前工作路径。

2）“Edit”（编辑）菜单（如图 1-3 所示）。编辑菜单除了具有 Windows 一般应用程序所具有的“撤销”、“重复”、“复制”、“粘贴”、“全选”外，还包括如下菜单项：

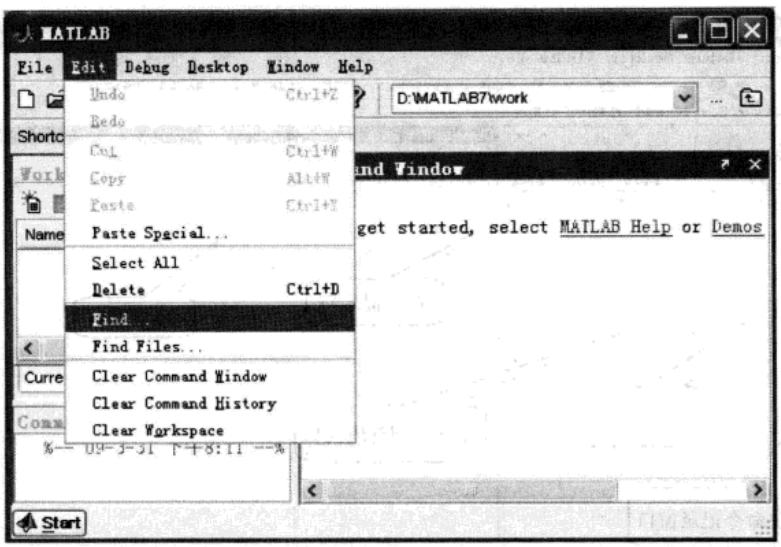


图 1-3 “Edit”（编辑）下拉菜单

“Clear Command Window”清除命令窗口；

“Clear Command History”清除命令的历史记录；

“Clear Workspace”清除工作空间。

3）“Debug”（调试）菜单（如图 1-4 所示）。Debug 菜单的各菜单项用于调试程序，具体功能如下：

图 1-4 “Debug”（调试）下拉菜单

“Open M-Files when Debugging”调试时打开文件；

“Step”调试时单步运行；

“Step In”调试时单步运行进入子函数；

“Step Out”调试时单步运行跳出子函数；

“Continue”运行程序到下一个断点或到程序结束；

“Clear Breakpoints in All Files”清除所有的断点；

“Stop if Errors/Warnings” 在程序出错或报警处停止；

“Exit Debug Mode” 退出调试模式。

4）“Desktop”（桌面）菜单（如图 1-5 所示）。为了改动 MATLAB 工作环境的外观，桌面菜单可以决定是否显示界面上摆布的一些窗口（界面布局）。读者可以对其中的每个菜单操作一下，看分别出现什么效果。

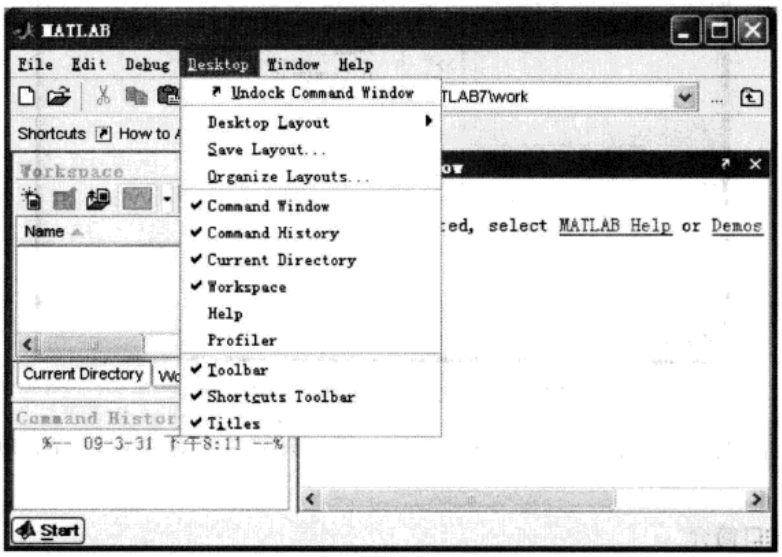


图 1-5　“Desktop”（桌面）下拉菜单

5）“Window”（窗口）菜单（如图 1-6 所示）。窗口菜单用于显示当前打开的 M 文件的文件名以及在已打开的窗口之间进行切换。

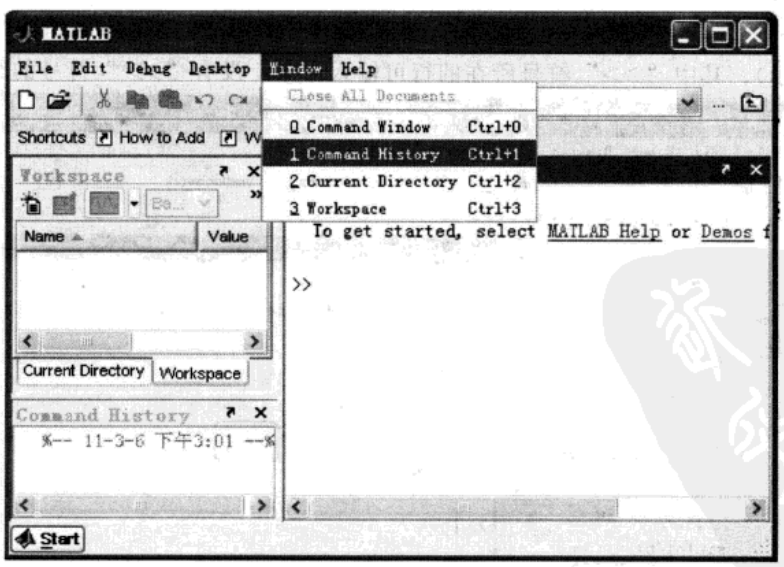


图 1-6　“Window”（窗口）下拉菜单

6）“Help”（帮助）菜单（如图 1-7 所示）。帮助菜单能为用户提供进入各类帮助系统的方法，通过菜单项打开帮助窗口，将显示各部分所需要的帮助信息。

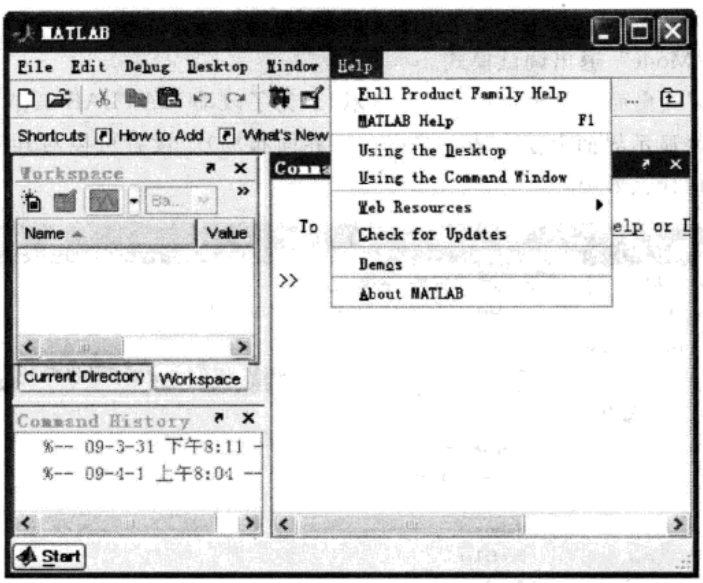


图 1-7 “Help”（帮助）下拉菜单

退出 MATLAB 和退出一般 Windows 程序一样，只需单击命令窗口右上角的“关闭”按钮即可。

2. 常用窗口简介

（1）Command Window（命令窗口）

命令窗口是对 MATLAB 进行操作的主要载体，默认情况下，启动 MATLAB 时就会打开命令窗口，显示形式如图 1-1 所示。一般来说，MATLAB 的所有函数和命令都可以在命令窗口中执行。在 MATLAB 命令窗口中，命令不仅可以由菜单操作来实现，也可以由命令行操作来执行。

例如，在命令窗口中输入 sin（pi/5），然后按【Enter】键，就会得到输出 ans＝0.5878（如图 1-8 所示）。其中“>>”符号所在的行可输入命令，没有“>>”符号的行显示结果。

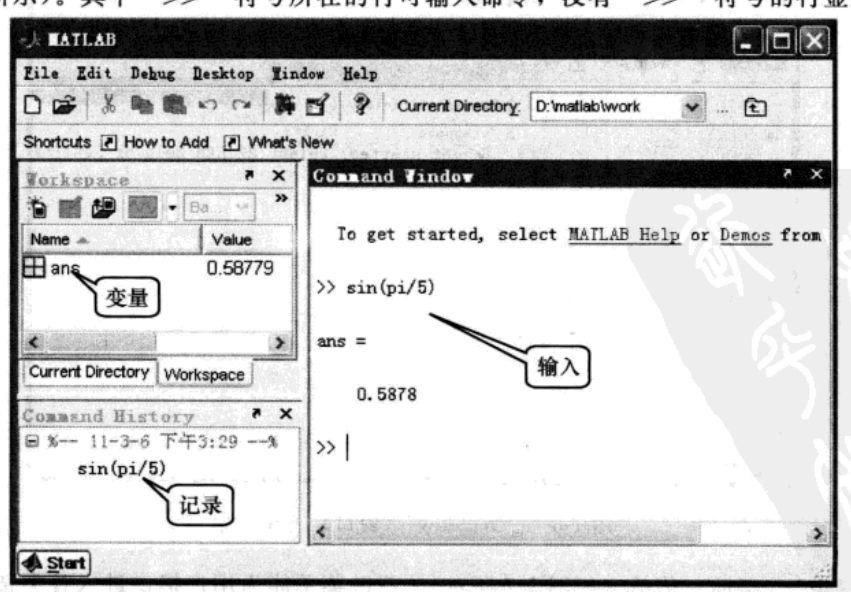


图 1-8 在命令窗口中输入命令

注意　在 MATLAB 命令行操作中，有一些键盘按键可以提供特殊而方便的编辑操作。比如【↑】可调出前一个命令行，【↓】可调出后一个命令行，这样避免了重新输入的麻烦。下面讲到的历史命令记录窗口也具有此功能。

(2) Command History（历史命令记录窗口）

该窗口记录着用户每一次开启 MATLAB 的时间，以及每一次开启 MATLAB 后，在 MATLAB 命令窗口中运行过的所有命令行（如图 1-7 所示）。这些命令行记录可以被复制到命令窗口中再运行，从而减少了重新输入的麻烦。选中该窗口中的任一命令记录，然后右击，则可根据菜单进行相应操作。或者双击某一行命令，也可在命令窗口中执行该命令（如图 1-9 所示）。

(3) Workspace（工作空间管理窗口）

在工作空间管理窗口中将显示所有目前保存在内存中的 MATLAB 变量的变量名及其对应的数据结构、字节数以及类型，而不同的变量类型分别对应不同的变量名图标（如图 1-8 所示）。选中一个变量，右击则可根据菜单进行相应的操作（如图 1-10 所示）。

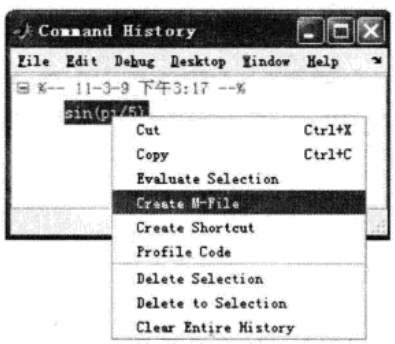


图 1-9　历史命令窗口的操作

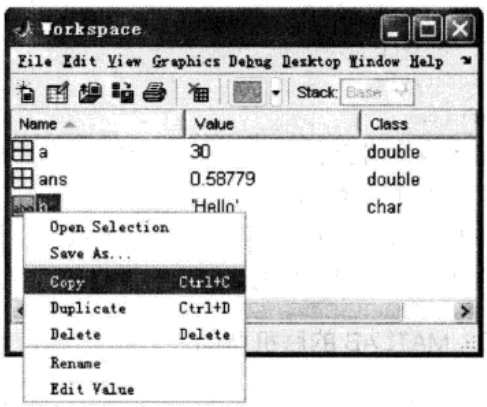


图 1-10　工作空间管理窗口的操作

(4) Current Directory（当前目录窗口）

在当前目录窗口中可显示或改变当前目录，还可以显示当前目录下的文件，包括文件名、文件类型、最后修改时间以及该文件的说明信息等并提供搜索功能（如图 1-11 所示）。

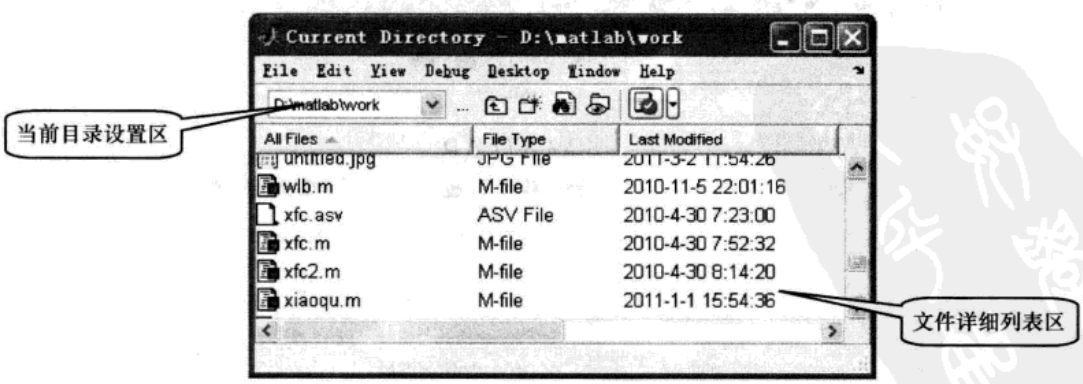


图 1-11　当前目录窗口

MATLAB 只执行当前目录或搜索路径下的命令、函数与文件。当前目录是指 MATLAB 运行文件时的工作目录，在当前目录窗口中可以显示或改变当前目录，还可以显示当前目录

下的文件并进行搜索。当用户在 MATLAB 命令窗口输入一条命令后，MATLAB 按照一定次序寻找相关的文件。基本的搜索过程是：1）检查该命令是不是一个变量；2）检查该命令是不是一个内部函数；3）检查该命令是否为当前目录下的 M 文件；4）检查该命令是否是 MATLAB 搜索路径中其他目录下的 M 文件。

用户可以将自己的工作目录列入 MATLAB 搜索路径，从而将用户目录纳入 MATLAB 系统统一管理。用对话框设置搜索路径的操作过程是：在 MATLAB 的“File”菜单中选择“Set Path”或在命令窗口执行“pathtool”命令，将出现搜索路径设置对话框。通过“Add Folder”或“Add with Subfolder”按钮将指定路径添加到搜索路径列表中。在修改完搜索路径后，需要将其保存。

（5）Figure（图形窗口）

在命令窗口输入 figure，可产生一个与命令窗口隔离的图形窗口。如在命令窗口输入如下命令：

```
t=0 :pi/100:2*pi;
y=sin(t);
plot(t,y)
grid on
```

plot 函数则会在图形窗口中绘制正弦曲线图形，如图 1-12 所示。

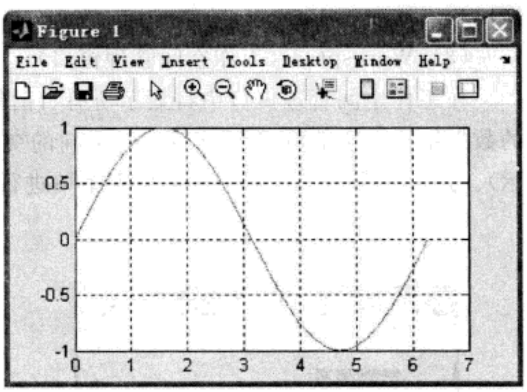


图 1-12 图形窗口

图形窗口和其他 Windows 窗口类似，有菜单栏与工具栏，能实现图形的编辑、修饰、存储等功能。

1.2.3 MATLAB 的联机帮助

MATLAB 和其他高级语言一样，具有完善的帮助系统。MATLAB 提供了相当丰富的帮助信息，同时也提供了获得帮助的方法。首先，可以通过桌面平台的“Help”菜单来获得帮助，也可以通过工具栏的帮助选项获得帮助。如在“Help”菜单下选择“MATLAB Help”项（如图 1-7 所示），则进入如图 1-13 所示的帮助导航窗口，在该窗口中可按需要查询一切命令的帮助信息。

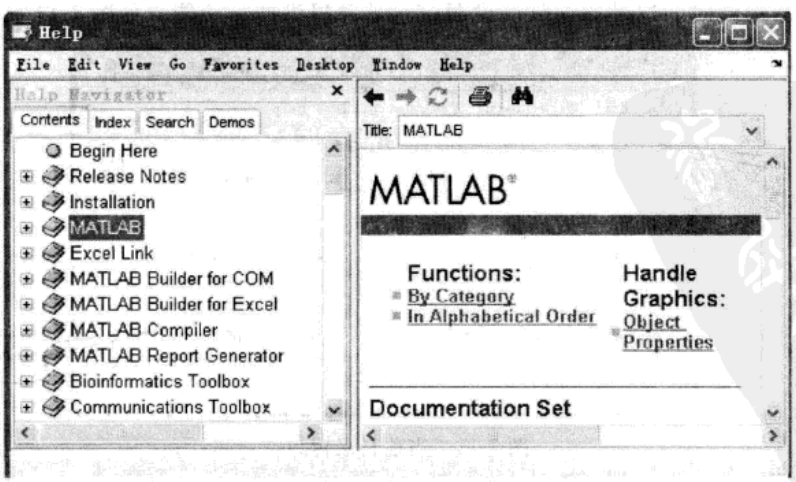


图 1-13 帮助导航窗口

在命令窗口中获得 MATLAB 帮助的命令及说明见表 1-1，其调用格式为

命令+ 指定参数

例如，在命令窗口中输入 help sin，即得到“sin”的相关信息和更多的用法介绍。

```
>> help sin
SIN       Sine.
SIN(X) is the sine of the elements of X.
See also asin, sind.
Overloaded functions or methods (ones with the same name in other directories)
help sym/sin.m
Reference page in Help browser
doc sin
```

表 1-1　命令窗口中获得 MATLAB 帮助的命令

命　令	说　明
help	在命令窗口中显示 M 文件的帮助
lookfor	在命令窗口中显示具有指定参数特征函数的 M 文件的帮助
doc	在帮助导航窗口中显示指定函数的参考信息
helpwin	打开帮助导航窗口，并且将初始界面置于 MATLAB 函数的 M 文件的帮助信息
helpdesk	打开一个名为“help”的帮助导航窗口
demo	打开一个“help”的演示模型界面，从而方便地了解 MATLAB 的基本功能

另外，MATLAB 7.0 支持模糊查询，用户只需要输入命令的前几个字母，然后按【Tab】键，系统就会列出所有以这几个字母开头的命令。

1.3　变量与函数

1.3.1　常量与变量

MATLAB 的数据类型主要包括数字、字符串、矩阵、单元型数据及结构型数据等。限于篇幅，本书将重点介绍其中几个常用类型。

1. 常量

MATLAB 中的数据有常量与变量之分，常量也称为数值。数值量包括实数和复数，具体形式上包括标量、向量、数组和矩阵等一切可以用数字表示的量。例如，实数一般采用十进制表示，可以带小数点和正负号，下面的数值都是合法的。

5、+5、−5.55、0.005 6、6.5e−5、100e60、−0.060e−0123

可以对数值量进行各种算术运算、关系运算和逻辑运算。

MATLAB 的计算都是以双精度（double）格式进行的，且所有数值量在内存中也都是以双精度保存的，但其显示格式却有不同形式，通常用户可在命令窗口中用 format 命令临时改变显示方式。比如用户希望以有理数（rational）形式显示，则可在命令窗口中输入命令“format rational”，如图 1-14 所示，数“0.25”的有理数显示形式为“1/4”。

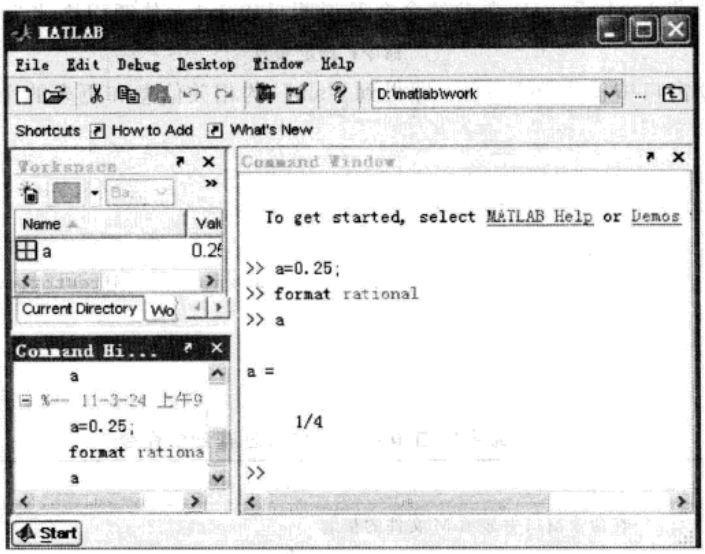


图 1-14　数值的显示格式

其他常用格式还有短格式（short，默认格式）和长格式（long），更多格式参见表 1-2。

表 1-2　数据的输出格式控制

格　式	中文解释	说　明
format	短格式（默认格式）	默认为短格式方式，与 format short 相同
format short	短格式	显示 5 位定点十进制数
format long	长格式	显示 15 位定点十进制数
format short e	短格式 e 方式	显示 5 位浮点十进制数
format long e	长格式 e 方式	显示 15 位浮点十进制数
format short g	短格式 g 方式	显示 5 位定点或 5 位浮点十进制数
format long g	长格式 g 方式	显示 15 位定点或 15 位浮点十进制数
format hex	十六进制格式	以十六进制格式显示
format +	+格式	以+、-和空格分别表示矩阵中的正数、负数和零元素
format bank	银行格式	按元、角、分（小数点后具有两位）的固定格式显示
format rat	有理数格式	用有理数逼近显示数据
format compact	压缩格式	数据之间无空行
format loose	自由格式	数据之间有空行

读者可自己在命令窗口中输入

```
x=1.2345e-6
```

然后在不同的输出格式下输出 x 的结果，观察结果显示的不同。

2. 变量

MATLAB 中的变量可用来存放数据，也可用来存放向量或矩阵，并进行各种运算。

变量的命名规则是：1）变量名区分大小写；2）变量名以字母开头，可以由字母、数字、下划线组成，但不能使用标点；3）变量名长度不超过 63 位，最多只能含有 63 个字符，后面的字符无效。

为了阅读程序的方便，对变量可作注释，“%”是注释符，“%”后面的内容为注释，对 MATLAB 的计算不产生任何影响。

MATLAB 将所有变量均保存为 double 的形式，在“Command Window”的状态下，所有的变量均存在于工作空间中。

3. 永久变量

永久变量是变量的一种特殊情况，它在工作空间中看不到，但是使用者可直接调用。表 1-3 列出了永久变量。

表 1-3　永久变量表

名称	取　值	名称	取　值
ans	用于结果的默认变量名	i，j	虚数单位：i=j=$\sqrt{-1}$
pi	圆周率 π 的近似值（3.141 6）	realmax	系统所能表示的最大数值
eps	数学中无穷小（epsilon）的近似值（2.220 4e－016）	realmin	系统所能表示的最小数值
inf	无穷大，如 1/0＝inf（infinity）	nargin	函数的输入参数个数
NaN	非数，如 0/0＝NaN（Not a Number），inf/inf＝NaN	nargout	函数的输出参数个数

在 MATLAB 中，定义变量时应避免与常量名重复，以免改变这些常量的值，如果已改变了某个常量的值，可以通过“clear＋常量名”命令恢复该常量的初始设定值（当然，也可通过重新启动 MATLAB 系统来恢复这些常量值）。

4. 符号变量

在 MATLAB 中进行符号运算时需要先用 syms 命令创建符号变量和表达式，如：

```
>> syms x
```

syms 不仅可以声明一个变量，还可以指定这个变量的数学特性，比如：

声明变量 x、y 为实数类型，可用命令：`>> syms x y real`

声明变量 x、y 为整数类型，可用命令：`>> syms x y positive`

5. 变量的查询与清除

在命令窗口中，只要输入“who”，就可以看到工作空间中所有曾经设定并至今有效的变量。如果输入“whos”，不但会显示所有的变量，而且会将该变量的名称、性质等都显示出来，即显示变量的详细资料。输入“clear”，就清除工作空间中的所有变量。如果输入“clear 变量名”，只清除工作空间中指定变量名的变量。

1.3.2　函数

MATLAB 系统提供了近 20 类基本命令函数，它们有一部分是 MATLAB 的内部命令，有一部分是以 M 文件形式出现的函数。这些 M 文件形式的函数扩展了 MATLAB 的功能，对于这些命令函数可以通过在命令行里面输入

```
Help fun
```

来获得有关这个命令函数使用的详细说明，这里 fun 是要查询的命令函数的名字。表 1-4 列出了基本的数学函数。

表 1-4　基本数学函数表

函　数　名	中 文 解 释	函　数　名	中 文 解 释
sin(x)	正弦函数	asin(x)	反正弦函数
cos(x)	余弦函数	acos(x)	反余弦函数
tan(x)	正切函数	atan(x)	反正切函数
exp(x)	以 e 为底的指数	log10(x)	以 10 为底数的对数
log(x)	自然对数	sqrt(x)	开平方
abs(x)	绝对值或向量的长度	max(x)	最大值
min(x)	最小值	sum(x)	元素求和
sign(x)	符号函数	round(x)	四舍五入到最近的整数
ceil(x)	朝正无穷方向取整	floor(x)	朝负无穷方向取整
fix(x)	朝零方向取整	gcd(x，y)	求两个整数的最大公约数

数学函数都有一个共同的特点：若自变量 x 为矩阵，则函数值也为 x 的同阶矩阵，即对 x 的每一元素分别求函数值；若自变量 x 为通常情况下的一个数值，则函数值是对应于 x 的一个数值。如图 1-15 所示为“sin(x)”的一个函数值与一组函数值的计算。

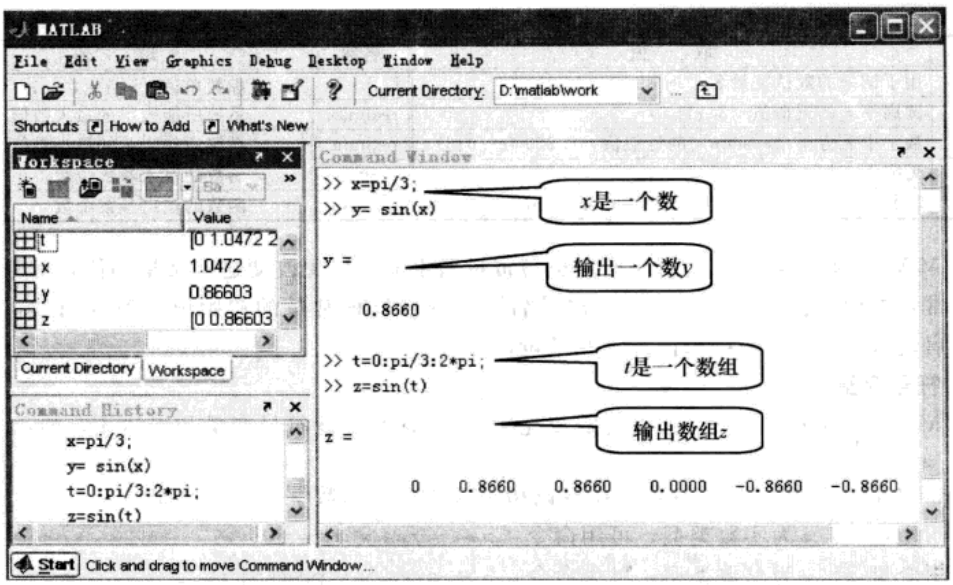


图 1-15　函数值的计算

1.4　矩阵及其运算

矩阵是 MATLAB 数据存储的基本单元，而矩阵的运算是 MATLAB 语言的核心，在 MATLAB 语言系统中几乎一切运算均是以对矩阵的操作为基础的。矩阵的运算是按一定的运算规则进行的，其规则是由运算符决定的。

1.4.1　操作符与运算符

1. 操作符

在编辑程序或命令中，当标点或其他符号表示特定的操作功能时就称其为操作符。表 1-5 列出了操作符。

表 1-5　操作符

操　作　符	使 用 说 明
:	冒号。①m：n 产生一个数组 $[m,m+1,\cdots,n]$；②m：k：n 产生一个数组 $[m,m+k,\cdots,n]$；③$A(:,j)$ 取矩阵 A 的第 j 列；④$A(k,:)$ 取矩阵 A 的第 k 行
;	分号。①在矩阵定义中表示一行的结束；②在命令语句的结尾表示不显示这行语句的执行结果
...	连续点。一个命令语句非常长一行写不完，可以分几行写，此时在行的末尾加上连续点，表示是一个命令语句
%	百分号。在编程时引导注释行，而系统解释执行程序时，%后面的内容不作处理

2. 运算符

算术运算符是构成运算的最基本的操作命令，可以在 MATLAB 的命令窗口中直接运行。运算符可分为三类：算术运算符、关系运算符与逻辑运算符。不同的运算符及功能说明见表 1-6、表 1-7、表 1-8。

表 1-6　算术运算符

运　算　符	功 能 说 明
+	加法运算。两个数相加或两个同阶矩阵相加。如果是一个矩阵和一个数字相加，则这个数字自动扩展为与矩阵同维的一个矩阵
−	减法运算。两个数相减或两个同阶矩阵相减
*	乘法运算。两个数相乘或两个可乘矩阵相乘
/	除法运算。两个数相除或两个可除矩阵相除（A/B 表示 A 乘以 B 的逆）
∧	乘幂运算。数的方幂或一个方阵的多少次方
\	左除运算。两个数 $a\backslash b$ 表示 $b\div a$，两个可除矩阵相除（$A\backslash B$ 表示 B 乘以 A 的逆）
. *	点乘运算。两个同阶矩阵对应元素相乘
./	点除运算。两个同阶矩阵对应元素相除
. ∧	点乘幂运算。一个矩阵中各个元素的多少次方
. \	点左除运算。两个同阶矩阵对应元素左除

表 1-7　关系运算符

运　算　符	功 能 说 明	运　算　符	功 能 说 明
>	判断大于关系	>=	判断大于等于关系
<	判断小于关系	<=	判断小于等于关系
==	判断等于关系	~=	判断不等于关系

表 1-8　逻辑运算符

运　算　符	功 能 说 明	运　算　符	功 能 说 明
&	与运算	~	非运算
|	或运算	xor（a，b）	异或运算

关系运算符主要用于比较数、字符串、矩阵之间的大小或不等关系，其返回值是 0 或 1。

逻辑运算符主要用于逻辑表达式和进行逻辑运算，参与运算的逻辑量以 0 代表“假”，以任意非 0 数代表“真”。逻辑表达式和逻辑函数的值以 0 表示“假”，以 1 表示“真”。

1.4.2　矩阵的输入与运算

1. 矩阵的输入

（1）直接输入法

从键盘直接输入矩阵的每一个元素。具体方法如下：将矩阵的所有元素用方括号括起来，在方括号内按矩阵行的顺序输入各元素，同一行的各元素之间用空格或逗号分隔，不同行的元素之间用分号或回车键分隔。例如：

```
>> A=[2,3,5;1,3,5;6,9,4]    % 同一行元素之间用空格或逗号分隔,行之间用分号或回车键分隔
A=
    2    3    5
    1    3    5
    6    9    4
```

(2) 外部文件读入法

MATLAB语言允许用户调用在MATLAB环境之外定义的矩阵。可以利用任意的文本编辑器编辑所要使用的矩阵，矩阵元素之间以特定分隔符分开，并按行列布置。load函数用于调用数据文件，其调用方法为：load+文件名［参数］。

例如：事先在记事本中编辑以下数据，保存为文件data1.txt，文件放在当前目录下。

```
1    1    1
1    2    3
1    3    6
```

在MATLAB命令窗口中输入：

```
>> load  data1.txt
>> data1                    % 显示数据
data1=
      1    1    1
      1    2    3
      1    3    6
```

load函数将会从文件名所指定的文件中读取数据，并将输入的数据赋给以文件名命名的变量，如果不给定文件名，则系统将自动认为MATLAB.mat文件为操作对象，如果该文件在MATLAB搜索路径中不存在，系统将会报错。

(3) 复制粘贴法

首先在命令窗口中输入矩阵名等于空的方括号（注意不要按回车键），如：

```
A=[];
```

其次，打开数据文件（如WORD、EXCEL形式的文件），复制文件中的数据；然后，返回命令窗口，将光标置于方括号内，右击，在弹出的快捷菜单中选择“粘贴”，这样数据就输入了。

读者若对Microsoft Excel有一定的使用经验。可使用MATLAB Excel Builder实现MATLAB和Microsoft Excel的连接，从而实现两者数据的无缝连接，更详细的操作请参考有关文献。

2. 特殊矩阵的建立

对于一些比较特殊的矩阵（如单位矩阵或元素中含1或0较多的矩阵），由于其具有特殊的结构，MATLAB提供了一些函数用于生成这些矩阵，见表1-9。

表1-9 生成特殊矩阵的命令函数

命令函数	功能说明
a=[]	生成空矩阵，当对一项操作无结果时，返回空矩阵，空矩阵的大小为零
b=zeros (m,n)	生成一个m行、n列的零矩阵
c=ones (m,n)	生成一个m行、n列的元素全为1的矩阵
d=eye (m,n)	生成一个m行、n列的单位矩阵
rand (m)	生成m阶均匀分布的随机矩阵
randn (m)	生成m阶正态分布的随机矩阵

3. 矩阵中元素或块的操作

对矩阵中元素或块的常用操作见表1-10。

表 1-10　矩阵中元素或块的常用操作

表达式或命令函数	功 能 说 明
$A(k,:)$	提取矩阵 A 的第 k 行
$A(:,k)$	提取矩阵 A 的第 k 列
$A(:)$	依次提取矩阵 A 的每一列，将 A 拉伸为一个列向量
$A(i1:i2,j1:j2)$	提取矩阵 A 的第 $i1$～$i2$ 行、第 $j1$～$j2$ 列，构成新矩阵
$A([a\ b\ c\ d],:)$	提取矩阵 A 的指定的第 a、b、c、d 行，构成新矩阵
$A(:,[e\ f\ g\ h])$	提取矩阵 A 的指定的第 e、f、g、h 列，构成新矩阵
$A(i2:-1:i1,:)$	以逆序提取矩阵 A 的第 $i1$～$i2$ 行，构成新矩阵
$A(:,j2:-1:j1)$	以逆序提取矩阵 A 的第 $j1$～$j2$ 列，构成新矩阵
$A(i1:i2,:)=[]$	删除 A 的第 $i1$～$i2$ 行，构成新矩阵
$A(:,j1:j2)=[]$	删除 A 的第 $j1$～$j2$ 列，构成新矩阵
$[A\ \ B]$ 或 $[A;B]$	将矩阵 A 和 B 拼接成新矩阵
diag(A, k)	抽取矩阵 A 的第 k 条对角线元素向量
tril(A, k)	抽取矩阵 A 的第 k 条对角线下面的部分
triu(A, k)	抽取矩阵 A 的第 k 条对角线上面的部分
flipud(A)	矩阵 A 进行上下翻转
fliplr(A)	矩阵 A 进行左右翻转
A'	矩阵 A 的转置
rot90(A)	矩阵 A 逆时针旋转 90°

例如：

```
>> A=[1,2,3;4,5,6;7,8,9]
A=
   1    2    3
   4    5    6
   7    8    9
>> A(2,:)                    % 取出 A 的第 2 行的所有元素
ans =
     4    5    6
>> A([1 3],[2,3])            % 取出 A 的 1,3 行与 2,3 列交叉的元素
ans =
     2    3
     8    9
```

4. 矩阵的运算

(1) 矩阵间的运算

矩阵间的运算见表 1-11。

表 1-11　矩阵间的运算

表　达　式	功 能 说 明
$A+B(A-B)$	A 与 B 为同型矩阵，对应元素相加（减）
$A*B$	A 的列数要等于 B 的行数，按代数学中定义的矩阵乘法法则计算
A/B	$X=A/B$ 是线性方程组 $XA=B$ 的解。当 A 是可逆的矩阵时，$A/B=A*B^{-1}$
$A\backslash B$	$X=A\backslash B$ 是线性方程组 $AX=B$ 的解。当 A 是可逆的矩阵时，$A\backslash B=A^{-1}*B$
$A.*B$	A 与 B 为同型矩阵，对应元素相乘
$A./B$	A 与 B 为同型矩阵，对应元素相除
$A.\wedge B$	A 与 B 为同型矩阵，A 中元素对应 B 中元素乘方运算

(2) 矩阵与标量的运算

矩阵与标量的运算见表 1-12。

表 1-12 矩阵与标量的运算

表 达 式	功能说明（设 A 为矩阵，c 为标量）
$A+c$ $(A-c)$	A 中每个元素加（减）常数 c
$A*c$ $(c*A)$	A 中每个元素乘常数 c
A/c	A 中每个元素除常数 c
$c./A$	常数 c 分别被 A 中对应每个元素相除
$c.\wedge A$	常数 c 与 A 中对应每个元素的乘方运算
$A.\wedge c$	对应与 A 中每个元素对应常数的 c 次乘方运算
$A\wedge c$	A 是方阵，当 c 大于 0 时表示矩阵的方幂，当 c 小于 0 时表示 A 逆的方幂

(3) 矩阵的基本函数运算

矩阵的函数运算是矩阵运算中最实用的部分，常用的主要有以下几个，见表 1-13。

表 1-13 矩阵的函数运算命令

命 令	功 能	命 令	功 能
det(A)	求矩阵 A 的行列式	rref(A)	求矩阵 A 的阶梯形的行最简形式
inv(A)	求方阵 A 的逆矩阵	rank(A)	求矩阵 A 的秩
size(A)	求矩阵 A 的阶数	trace(A)	求矩阵 A 的迹
eig(A)	求 A 的特征值及特征向量	$[Q, R]=qr(A)$	求正交矩阵 Q 和上三角阵 R 满足 $A=QR$
orth(A)	将非奇异矩阵 A 正交规范化		

(4) 矩阵的数据处理

MATLAB 具有强大的数据处理功能，比如数据的排序、求最大值、求和、求均值等。本书第 5 章将详细地介绍一些数据分析方法。常用数据处理的命令见表 1-14。

表 1-14 常用数据处理的命令

命 令	功 能	命 令	功 能
max(A)	求向量或矩阵列的最大值	min(A)	求向量或矩阵列的最小值
mean(A)	求向量或矩阵列的平均值	median(A)	求向量或矩阵列的中间值
sum(A)	求向量或矩阵列的元素和	prod(A)	求向量或矩阵列的元素乘积
var(A)	求向量或矩阵列的方差	std(A)	求向量或矩阵列的标准差
cov(A)	矩阵列向量之间的协方差矩阵	corrcoef(A)	矩阵列向量之间的相关系数矩阵
length(A)	求向量所含元素个数	find(A)	求向量中满足条件的元素

1.4.3 数组的输入与运算

只有一行的矩阵也称为数组或向量，MATLAB 中对数组设置了一些相对于矩阵不一样的创建或运算命令。

1. 数组的输入

表 1-15 创建简单数组的方法

命 令	用 途
$x=[a,b,c,d]$	创建包含指定元素的行向量
$x=$first:last	创建从 first 开始，加 1 计数，到 last 结束的行向量
$x=$first:increment:last	创建从 first 开始，加 increment 计数，以 last 结束的行向量

（续）

命　令	用　途
x=linspace(first,last,n)	创建从 first 开始，到 last 结束，有 n 个元素的行向量
x=logspace(first,last,n)	创建从 first 开始，到 last 结束，有 n 个元素的对数分隔行向量
x=[y，z，1，2，3]	y 和 z 都是行向量

2. 数组元素的访问

访问一个元素：x（i）表示访问数组 x 的第 i 个元素。

访问一块元素：$x(s:h:t)$ 表示访问数组 x 的从第 s 个元素开始，以步长 h 到第 t 个（但不超过 t）的这些元素，h 可以为负数，h 默认为 1。

直接使用元素编址序号：$x([a，b，c，d])$ 表示提取数组 x 的第 a、b、c、d 个元素构成一个新的数组 $[x(a)\quad x(b)\quad x(c)\quad x(d)]$。

3. 标量与数组的运算

标量与数组的加、减、乘、除、乘方运算是数组的每个元素与该标量施加相应的加、减、乘、除、乘方运算，其表达式见表 1-16。

表 1-16　标量与数组的运算

表　达　式	运　算　结　果
$a+c$	$=[a1+c,a2+c,\cdots,an+c]$，即数组 a 的每个元素加上 c
$a*c$ 或 $a.*c$	$=[a1*c,a2*c,\cdots,an*c]$，即数组 a 的每个元素乘以 c
a/c 或 $a./c$	$=[a1/c,a2/c,\cdots,an/c]$，即数组 a 的每个元素除以 c
$a.\backslash c$	$=[c/a1,c/a2,\cdots,c/an]$，即 c 除以数组 a 的每个元素
$a.\wedge c$	$=[a1\wedge c,a2\wedge c,\cdots,an\wedge c]$，即数组 a 的每个元素的 c 次幂
$c.\wedge a$	$=[c\wedge a1,c\wedge a2,\cdots,c\wedge an]$，即以 c 为底，以 a 的每个元素为指数的幂

其中 $a=[a1,a2,\cdots,an]$是数组，c 为标量。

4. 数组与数组的运算

数组与数组的运算要求数组维数是相同的，其加、减、乘、除、幂运算可按元素对元素方式进行，不同维数的数组不能进行运算，其表达式见表 1-17。

表 1-17　数组与数组的运算

表　达　式	运　算　结　果
$a+b$	$=[a1+b1,a2+b2,\cdots,an+bn]$，即数组 a 与 b 的对应元素相加
$a.*b$	$=[a1*b1,a2*b2,\cdots,an*bn]$，即数组 a 与 b 的对应元素相乘
$a./b$	$=[a1/b1,a2/b2,\cdots,an/bn]$，即数组 a 与 b 的对应元素相除
$a.\backslash b$	$=[b1/a1,b2/a2,\cdots,bn/an]$，即数组 b 与 a 的对应元素相除
$a.\wedge b$	$=[a1\wedge b1,a2\wedge b2,\cdots,an\wedge bn]$，即数组 a 与 b 的对应元素的幂

其中 $a=[a1,a2,\cdots,an]$，$b=[b1,b2,\cdots,bn]$。

注意　数组的乘除法是指两个同维数组对应元素之间的乘除法，它们的运算符只能为“.*”、“./”或“.\”，而表达式 $a*b$、a/b、$a\wedge b$ 是没有意义的。

1.5　M 文件与编程

1.5.1　M 文件编辑/调试器窗口

在默认状态下，M 文件编辑/调试器窗口（Editor/Debugger）不随 MATLAB 界面的出现而启动。只有需要编写 M 文件时，才启动该窗口。如图 1-16 所示。

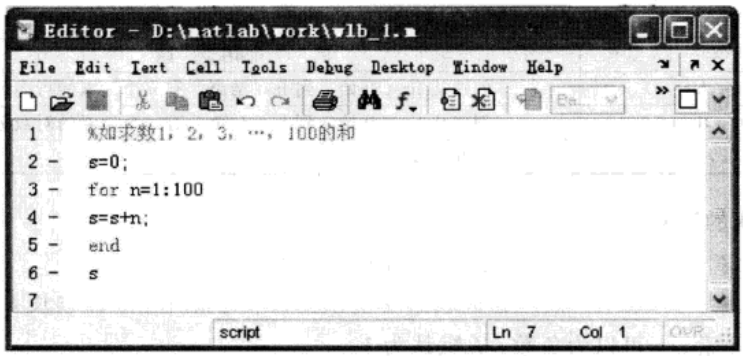


图 1-16 M 文件编辑/调试器窗口

M 文件编辑/调试器的菜单与工具栏请参考帮助。

启动该窗口有如下三种方法：1）菜单操作。选择 MATLAB 命令窗口的“File”菜单中的“New”菜单项单击“M-file 命令”。2）命令操作。在 MATLAB 命令窗口输入命令 edit。3）命令按钮操作。单击 MATLAB 命令窗口工具栏上的“新建”命令按钮。

1.5.2 M 文件

M 文件是由 MATLAB 语句（命令或函数）构成的 ASCII 码文本文件，文件名必须以“.m”为扩展名。M 文件通过 M 文件编辑/调试器生成。在命令窗口调用 M 文件，可实现一次执行多条 MATLAB 语句的功能。M 文件有以下两种形式。

1. 命令文件

命令文件是 MATLAB 命令或函数的组合，没有输入输出参数，执行命令文件只需在命令窗口中键入文件名按回车或在 M 文件编辑/调试器窗口激活状态下按“F5”键。

当用户要运行的指令较多时，可以直接从键盘上逐行输入指令，但这样做显得很麻烦，而命令文件则可以较好地解决这一问题。用户可以将一组相关命令编辑在同一个 ASCII 码命令中，运行时只需输入文件名，MATLAB 就会自动按顺序执行文件中的命令。这类似于批处理文件。命令文件中的语句可以访问 MATLAB 工作空间（Workspace）中的所有数据。在运行的过程中所产生的变量均是全局变量。这些变量一旦生成，就一直保存在内存空间中，除非用户将它们清除（如 clear 命令）。

如求数 1，2，3，4，…，100 的和。

在编辑器中写出程序如下：

```
s= 0;
for n= 1:100
s= s+ n;
end
s
```

保存为 wlb_1（这是文件名），然后在命令窗口中执行，即输入文件名：

```
>> wlb_1
s=
   5050              % 这是程序运行的结果
```

2. 函数文件

函数文件是另一种形式的 M 文件，可以有输入参数和返回输出参数，函数在自己的工作空间中操作局部变量，它的第一句可执行语句是以 function 引导的定义语句。在函数文件中的

变量都是局部变量，它们在函数执行过程中驻留在内存中，在函数执行结束时自动消失。函数文件不单单具有命令文件的功能，更重要的是它提供了与其他 MATLAB 函数和程序的接口，因此功能更强大。

MATLAB 函数文件的格式为：

```
function [返回参数 1,参数 2,…]= 函数名(输入参数 1,参数 2,…)
函数体
```

注意　第一行的有无，是区分命令文件与函数文件的重要标志；函数体包含所有函数程序代码，是函数的主体部分；函数文件保存的文件名应与用户定义的函数名一致。在命令窗口中以固定格式调用函数。

例如，定义函数 $f(x, y)=x^3+y^3-3xy$，并计算 $f(2, 3)$。

在编辑器中写出如下程序：

```
function  f= wlb_2(x,y)                    % 函数名为 wlb_2,返回值为 f
f= x.^3+ y.^3- 3*x.*y;                     % 这是函数主体
```

保存为 wlb_2（这是文件名，与函数名一致），然后在命令窗口中执行

```
>> wlb_2(2,3)
ans =
     17
```

1.5.3　控制语句的编程

1. 循环语句

MATLAB 提供了两种循环方式：for…end 循环和 while…end 循环。

1）for 循环语句（计数循环方式），其调用格式如下：

```
for 循环变量= 初值:步长:终值
      循环体
end
```

其执行过程为：将初值赋给循环变量，执行循环体；执行完一次循环之后，循环变量自增一个步长的值，然后再判断循环变量的值是否介于初值和终值之间，如果满足，仍然执行循环体，直至不满足为止。

2）while 循环语句（条件循环方式），其一般调用格式如下：

```
while 表达式
    循环体
end
```

其执行过程为：若表达式的值为真，则执行循环体语句，执行后再判断表达式的值是否为真，直到表达式的值为假时跳出循环。

while 语句一般用于事先不能确定循环次数的情况。

2. 条件控制语句

1）if...else...end 语句，其调用格式如下：

```
if 表达式
      语句体 1;
  else
      语句体 2;
end
```

其执行过程为：当表达式的值为真时，执行语句体 1，否则执行语句体 2；语句体 1 或语句体 2 执行后，再执行 if 语句的后继语句。

2）switch 分支结构语句，其调用格式如下：

```
switch  表达式
  case  表达式 1
          语句体 1
  case  表达式 2
          语句体 2
          …
  case  表达式 m
          语句体 m
  otherwise
          语句体 m+1
end
```

其执行过程为：控制表达式的值与每一个 case 后面表达式的值比较，若与第 k（k 的取值为 1～m）个 case 后面的表达式 k 的值相等，就执行语句体 k；若都不相同，则执行 otherwise 下的语句体 $m+1$。

3）其他流程控制语句，包括 continue 语句、break 语句和 return 语句。

- continue 语句用于 for 循环和 while 循环中，其作用就是终止一次循环的执行，跳过循环体中所有剩余的未被执行的语句，去执行下一次循环。
- break 语句也常用于 for 循环和 while 循环中，其作用就是终止当前循环的执行，跳出循环体，去执行循环体外的下一行语句。
- return 语句用于终止当前的命令序列，并返回到调用的函数或键盘，也用于终止 keyboard 方式。

1.6 MATLAB 通用操作实例

下面通过一个操作实例，说明 MATLAB 的通用操作界面的使用方法，使读者对软件环境更加熟悉，并且掌握如何在命令窗口中使用简单命令。

按照以下步骤进行。

1）启动 MATLAB。

2）在命令窗口中输入以下几行命令：

```
>> a=[1,2,3;4,5,6;7,8,9];
>> b=[1,3,5;2,4,6;5,7,9];
>> c='矩阵加法计算';
>> d=a+b;
>> wlb='矩阵乘法计算';
>> w=a*b;
```

3）打开工作空间管理窗口查看变量，共有 6 个变量，如图 1-17 所示。

4）双击其中的变量“a”，出现数组编辑器窗口（Array Editor），如图 1-18 所示为该变量的详细信息。

图 1-17　工作空间管理窗口中的 6 个变量

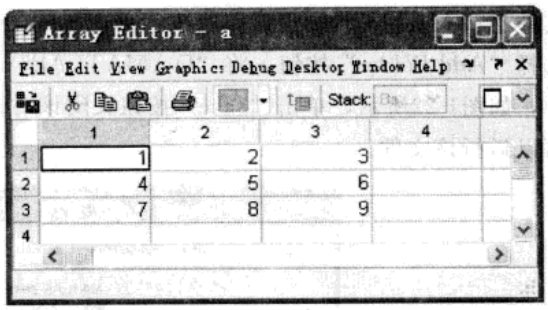


图 1-18　变量“a”的详细信息

5）打开历史命令记录窗口（Command History），用光标选中上面的 6 行命令，右击，在快捷菜单中选择“Create M-File”命令生成 M 文件，如图 1-19 所示。

6）在 M 文件编辑/调试器窗口（Editor/Debugger）（如图 1-20 所示）中，单击工具栏的“Save”按钮，将文件保存为“D:\MATLAB7\work\jinjishuxueshiyan\shiyan101.m”。

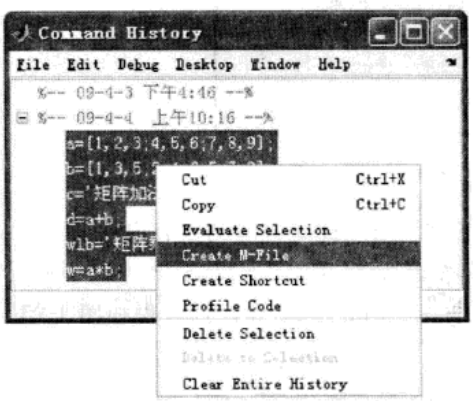


图 1-19　生成 M 文件

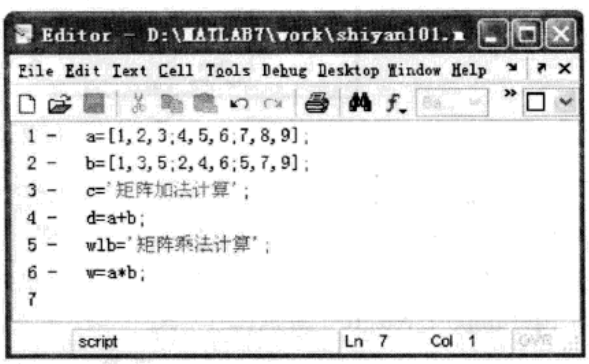


图 1-20　保存文件

7）打开当前目录窗口（Current Directory），将当前目录设置为“D:\MATLAB7\work\jinjishuxueshiyan”，可以看到刚保存的“shiyan101.m”文件，在命令窗口中输入“shiyan101”运行文件。

8）在命令窗口中输入“save shiyan101”命令，从当前目录窗口可以看到在当前目录下生成了一个“shiyan101.mat”数据文件，如图 1-21 所示。

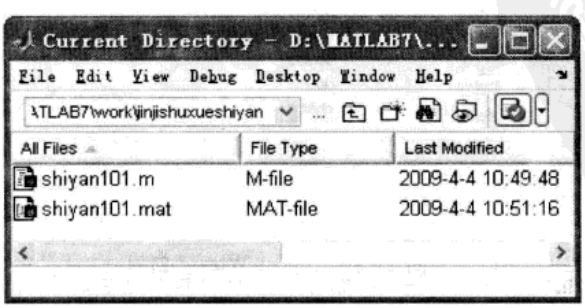


图 1-21　生成 MAT 文件

9）在命令窗口中输入“exit”命令退出 MATLAB。

10）重新启动 MATLAB，在命令窗口中输入“shiyan101”此时不能运行该文件，因为该文件不在 MATLAB 的搜索路径中。单击界面的菜单“File”→“Set Path”，打开设置对话框，选择“Add Folder”按钮，将“D:\MATLAB7\work\jinjishuxueshiyan”目录添加到搜索路径中，如图1-22所示，单击“save”按钮关闭该对话框，重新在命令窗口中输入“shiyan101”则可以运行该文件。

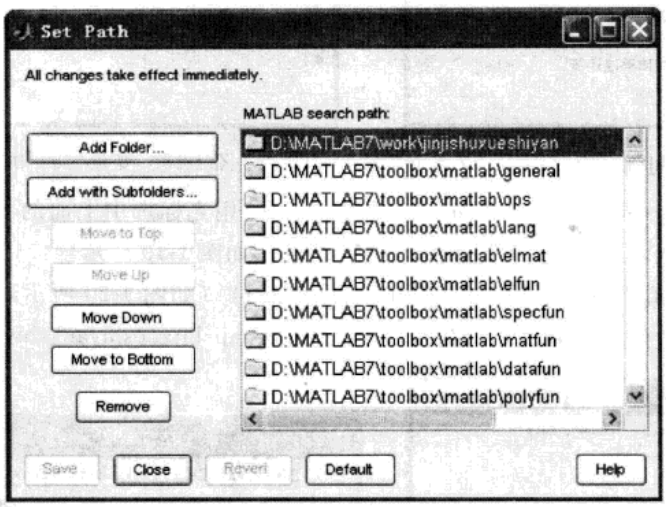


图 1-22　添加目录到搜索路径

11）退出 MATLAB 后重新启动，打开工作空间管理窗口，此时将看到没有内存变量。如果要将“shiyan101.mat”数据文件的变量导入，可选择菜单“File”→“Import Data”命令，然后选择“D:\MATLAB7\work\jinjishuxueshiyan\shiyan101.mat”文件，打开得如图 1-23 所示的“Import Wizard”窗口。

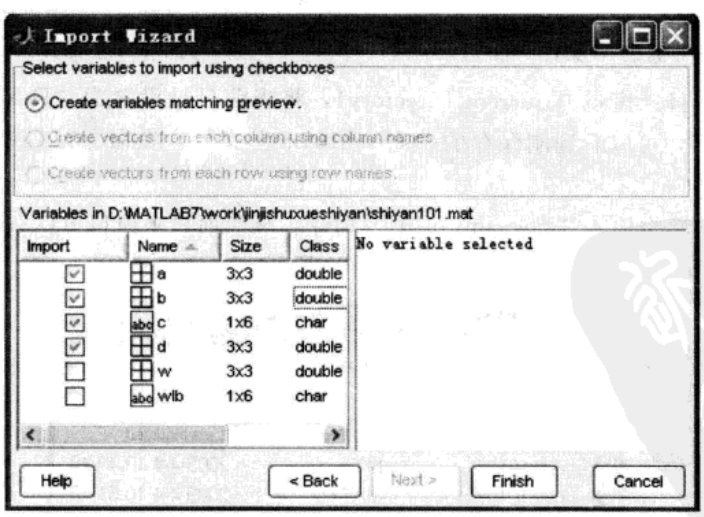


图 1-23　Import Wizard 窗口

在该窗口中将要导入的变量前的复选框选中，如选中“*a*”、“*b*”、“*c*”、“*d*”4 个变量，然后单击“finish”按钮，查看工作空间中出现的 4 个变量。

12）如果要查看文件“shiyan101.m”的内容，只要在 MATLAB 命令窗口中输入“type shiyan101”命令，则可看到文件的内容如下：

```
>> type shiyan101
a= [1,2,3;4,5,6;7,8,9];
b= [1,3,5;2,4,6;5,7,9];
c= '矩阵加法计算';
d= a+ b;
wlb= '矩阵乘法计算';
w= a * b;
```

习　题　1

1. 熟悉 MATLAB 7.0 的桌面平台的菜单栏和工具栏。
2. 分别使用直接输入法、外部文件读入法和复制粘贴法，来创建一个矩阵。
3. 先生成两个矩阵：A=[3 6 9 5；2 4 8 3；1 2 3 7；5 1 4 8] 和 B=[1 2 3 2；2 4 1 5；1 4 7 2；7 4 2 9]，后求解 $A.*B$、$A.\wedge B$、$A.\backslash B$ 和 $A./B$ 的结果。
4. 输入任意矩阵 A、B（它们的元素个数相等），命令 $A(:)$ 和 $A(:)=B$ 会产生什么结果？
5. 输入矩阵 A=[1，3，5；5，8，3；6，1，6]、B=[3，6；9，3；4，7]、C=[3，7，9，4，0，7]、D=2:6，体会命令 [A，B]、[A；C]、[A，B；D] 所产生的结果，总结由小矩阵生成大矩阵的方法。
6. 设 $f(x，y)=x^2+\sin xy+2y$，在 M 文件编辑/调试器中创建一个名为 wlb_3 的 M 函数文件并保存，在命令窗口中调用 M 文件，实现输入自变量的值时输出函数值。

第2章

数据描述性分析

数据描述性分析是从样本数据出发，概括分析数据的集中位置、分散程度、相互关联关系，以及数据分布的正态或偏态特征等。它是进行数据进一步分析的基础，对不同类型量纲的数据有时还要进行变换，然后再作出合理分析。本章主要介绍样本数据的基本统计量、数据的可视化、数据分布检验及数据变换等内容。

2.1 基本统计量与数据可视化

2.1.1 样本数据的基本统计量

描述数据基本特征主要为集中位置和分散程度。

设从所研究的对象（即总体）X 中观测得到 n 个观测值

$$x_1, x_2, \cdots, x_n \tag{2.1.1}$$

这 n 个值称为样本数据，简称数据，n 称为样本容量。

我们的任务就是要对式（2.1.1）进行分析，提取数据中所包含的有用信息，从而进一步对总体的特性作出推断。

1. 均值、中位数、分位数与三均值

式（2.1.1）的平均值称为样本均值，记为

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1.2}$$

样本均值描述了数据取值的平均位置。样本均值计算简易，但易受异常值的影响而不稳健。

将式（2.1.1）中的数据按从小到大的次序排列，排序为 k 的数记为 $x_{(k)}$（$1 \leqslant k \leqslant n$），即 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$，则

$$x_{(1)}, x_{(2)}, \cdots, x_{(n)} \tag{2.1.3}$$

称为样本数据的次序统计量。

由次序统计量定义数 M，

$$M = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ 为奇数} \\ \dfrac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & n \text{ 为偶数} \end{cases} \tag{2.1.4}$$

称 M 为样本数据的中位数。

中位数是描述数据中心位置的数字特征，比中位数大或小的数据个数大约为样本容量的一半。若数据的分布对称，则均值与中位数比较接近。若数据的分布为偏态，则均值与中位

数差异会较大。中位数的一个显著特点是受异常值的影响较小，具有较好的稳健性。

设 $0 \leqslant p < 1$，样本数据的 p 分位数定义为

$$M_p = \begin{cases} x_{([np]+1)} & np \text{ 不是整数} \\ \dfrac{1}{2}(x_{(np)} + x_{(np+1)}) & np \text{ 为整数} \end{cases} \tag{2.1.5}$$

其中 $[np]$ 表示 np 的整数部分。

显然，当 $p=0.5$ 时，$M_{0.5}=M$，即数据的 0.5 分位数等于其中位数。

一般来说，从整批数据（总体）中抽取样本数据，则整批数据中约有 $100p\%$ 个不超过样本数据的 p 分位数。在实际应用中，0.75 分位数与 0.25 分位数比较重要，它们分别称为上、下四分位数，记为 Q_3、Q_1。

一方面，虽然均值 $\bar{x}$ 与中位数 M 都是描述数据集中位置的数字特征，但 $\bar{x}$ 用了数据的全部信息，M 只用了部分信息，因此通常情况下均值比中位数有效；另一方面，当数据有异常值时，中位数比较稳健。为了兼顾两方面的优势，人们提出三均值的概念，并定义三均值如下：

$$\hat{M} = \frac{1}{4}M_{0.25} + \frac{1}{2}M + \frac{1}{4}M_{0.75} \tag{2.1.6}$$

由定义，三均值是上四分位数、中位数与下四分位数的加权平均，即分位数向量（$M_{0.25}$，M，$M_{0.75}$）与权向量（0.25，0.5，0.25）的内积。

在 MATLAB 中，提供了求均值、中位数、分位数等的命令函数。

1）均值命令 mean，其调用格式为：

```
m= mean(X);
```

其中，输入 X 为样本数据；输出 m 为样本均值。

2）中位数命令 median，其调用格式为：

```
MD= median(X);
```

其中，输入参数 X 是样本数据；输出 MD 为中位数。

3）P 分位数命令 prctile，其调用格式为：

```
SM= prctile(X,P);
```

其中，输入参数 X 是样本数据；P 为介于 0～100 的整数，$P=100*p$（$0 \leqslant p \leqslant 1$）；输出 SM 为 $P\%$ 分位数。

注意　当样本数据 X 是矩阵时，上述 3 个命令的输出将给出对应于 X 每列数据的数值，参见例 2.1.1。

4）根据分位数命令及式（2.1.6），可编写求三均值的 MATLAB 程序如下：

```
w= (0.25,0.5,0.25);                    % 输入权向量 w
SM= w* prctile(X,[25,50,75]);          % 由式(2.1.6)计算 X 三均值
```

例 2.1.1　表 2-1 是 2008 年安徽省各市森林资源情况统计数据，计算各指标均值、中位数以及三均值。

表 2-1　安徽省各市森林资源情况（2008 年）

地　区	林业用地面积（千公顷）	森林面积（千公顷）	森林覆盖率（%）	活立木总蓄积量（万立方米）	森林蓄积量（万立方米）
合肥市	53.93	50.98	15.48	256.00	65.41
淮北市	44.92	40.38	14.99	211.07	151.14
亳州市	148.19	145.54	17.10	842.09	677.52

（续）

地　区	林业用地面积（千公顷）	森林面积（千公顷）	森林覆盖率（%）	活立木总蓄积量（万立方米）	森林蓄积量（万立方米）
宿州市	293.86	279.86	28.80	1 238.01	1 035.67
蚌埠市	86.96	74.64	12.91	302.67	299.32
阜阳市	165.62	160.25	16.46	898.76	800.96
淮南市	17.93	16.37	6.20	151.39	30.17
滁州市	199.46	158.24	11.90	885.16	591.17
六安市	660.36	607.16	34.74	2 278.37	1 984.36
马鞍山市	17.14	13.72	8.10	81.20	36.34
巢湖市	148.52	117.54	12.60	494.38	335.26
芜湖市	77.27	66.69	20.85	279.34	187.92
宣城市	724.30	640.15	54.00	2 446.98	2 323.04
铜陵市	36.78	32.10	32.12	137.64	115.10
池州市	539.49	458.66	56.86	2 277.00	2 237.43
安庆市	598.92	546.67	35.60	2 291.09	2 099.21
黄山市	791.50	680.96	77.80	3 298.56	3 252.88

资料来源：《安徽统计年鉴 2009》。

解：按第 1 章介绍的矩阵输入方法，首先将表 2-1 中的数据作为矩阵 A 输入 MATLAB，然后对矩阵 A 调用有关命令函数，程序如下：

```
A= [53.93,…,3252.88];          % 输入数据,A 的每一列是表 2-1 对应指标的样本数据
M= mean(A);                    % 计算各指标(即各列)均值
MD= median(A);                 % 计算各指标中位数
SM= [0.25,0.5,0.25]*prctile(A,[25,50,75]);   % 计算各指标三均值
[M;MD;SM]                      % 输出计算结果(表 2-2)
```

表 2-2　安徽省森林资源均值、中位数与三均值（2008 年）

统　计　量	林业用地面积	森 林 面 积	森林覆盖率	活立木总蓄积量	森林蓄积量
均值	270.9	240.6	26.9	80.6	954.3
中位数	148.5	145.5	17.1	842.1	591.2
三均值	225.8	205.0	20.5	1 051.6	834.4

2. 方差与变异系数

方差是描述数据取值分散性的一种度量，它是数据相对于均值的偏差平方的平均。样本数据的方差记为

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n(\overline{x})^2\right) \tag{2.1.7}$$

其算术平方根称为标准差或根方差，即

$$s = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2 - n(\overline{x})^2\right)} \tag{2.1.8}$$

变异系数是描述数据相对分散性的统计量，其计算公式为

$$v = s/\overline{x}, \text{或者 } v = s/|\overline{x}| \tag{2.1.9}$$

变异系数是一个无量纲的量，一般用百分数表示。

在 MATLAB 中，计算方差的命令为 var，其调用格式为：

```
S= var(x);
```

计算标准差的命令为 std，其调用格式为：

```
d= std(x)
```

其中，输入 x 是样本数据；输出 S 为方差；d 为标准差。当输入 x 是矩阵时，输出 x 每列数据的方差与标准差。

由均值与方差命令，可设计变异系数的计算程序为：

```
v= std(x)./mean(x),或者 v= std(x)./abs(mean(x))
```

当输入 x 是矩阵时，输出 x 每列数据的变异系数。

例 2.1.2　计算例 2.1.1 中各指标的方差、标准差与变异系数。

解：将表 2-1 中的数据作为矩阵 A 输入，然后调用有关计算命令，程序如下：

```
A= [53.93,…,3252.88];          % 输入原始数据(注:为节约篇幅,大部分数据用省略号表示了)
M= mean(A);                    % 计算各指标均值
D= var(A);                     % 计算各指标方差
SD= std(A);                    % 计算各指标标准差
V= SD./abs(M)                  % 计算各指标变异系数
[D;SD;V]                       % 输出计算结果
```

结果见表 2-3。

表 2-3　安徽省森林资源方差、标准差与变异系数（2008 年）

统　计　量	林业用地面积	森 林 面 积	森林覆盖率	活立木总蓄积量	森林蓄积量
方差	75 464.48	59 198.14	394.49	1 065 554.98	1 040 590.73
标准差	274.71	243.31	19.86	1 032.26	1 020.09
变异系数	1.01	1.01	0.74	0.96	1.07

3. 样本的极差与四分位极差

样本数据的极大值与极小值的差称为**极差**，其计算公式为

$$R = x_{(n)} - x_{(1)} \tag{2.1.10}$$

极差是一种较简单的表示数据分散性的数字特征。

样本数据上、下四分位数 Q_3、Q_1 之差称为四分位极差，即

$$R_1 = Q_3 - Q_1 \tag{2.1.11}$$

四分位极差也是度量数据分散性的一个重要数字特征。由于分位数对异常值有抗扰性，所以四分位极差对异常数据也具有抗扰性。

在 MATLAB 中，求极差的命令为 range，其调用格式为：

```
R= range(x)
```

其中，输入 x 是样本数据；输出 R 是极差。求四分位极差的命令为 iqr，其调用格式为：

```
R1= iqr(x)
```

其中，输入 x 是样本数据；输出 $R1$ 是四分位极差。

4. 异常点判别

在解决实际问题时需要对异常数据进行处理。一般判别异常值比较简单的方法是：首先，计算数据的上截断点

$$Q_1 + 1.5R_1$$

与下截断点

$$Q_3 - 1.5R_1$$

其次，将数据逐个与截断点比较，小于下截断点的数据为特小值，大于上截断点的数据为特大值，两者均判为异常值。

例 2.1.3 根据 2007 年华东各地区高校教职工数据（表 2-4），计算专任教师、行政人员、教辅人员以及工勤人员占在职教工的百分比，并计算该百分比的极差、四分位极差及上、下截断点。

表 2-4 2007 年华东各地区高校教职工数据

地 区	在职教工	专任教师	行政人员	教辅人员	工勤人员
上海	61 385	35 480	10 282	7 842	7 781
江苏	134 215	88 568	20 172	13 371	12 104
浙江	67 763	45 622	10 960	6 798	4 383
安徽	59 149	40 743	7 278	5 763	5 365
福建	47 864	31 385	7 712	5 034	3 733
江西	63 392	45 153	8 179	5 495	4 565
山东	120 996	81 889	16 342	11 614	11 151

数据来源：《中国统计年鉴 2008》。

解：将表 2-4 中的数据作为矩阵 A 输入，然后调用有关计算命令，程序如下：

```
A= [61385     35480    10282    7842     7781
   134215     88568    20172    13371    12104
    67763     45622    10960    6798     4383
    59149     40743    7278     5763     5365
    47864     31385    7712     5034     3733
    63392     45153    8179     5495     4565
   120996     81889    16342    11614    11151];     % 输入表 2-4 中数据
B= A(:,2:5)./[A(:,1) * ones(1,4)];                    % 计算百分比
R= range(B);                                          % 计算百分比极差
R1= iqr(B);                                           % 计算四分位极差
XJ= prctile(B,[25])- 1.5 * R1;                        % 计算下截断点
SJ= prctile(B,[75])+ 1.5 * R1;                        % 计算上截断点
```

由程序的运行结果可知，上海市的专任教师占在职教工的百分比、教辅人员以及工勤人员占在职教工的百分比为异常值。

5. 偏度与峰度

数据分布特征一般用偏度与峰度描述。偏度是用于衡量分布的不对称程度或偏斜程度的指标。

样本数据的偏度定义为

$$p_d = \frac{n^2 u_3}{(n-1)(n-2)s^3} \tag{2.1.12}$$

其中，u_3、s 分别表示样本数据的 3 阶中心矩与标准差。

当 $p_d > 0$ 时称数据的分布是右偏的，此时均值右边的数据比均值左边的数据分布更散；当 $p_d < 0$ 时称数据的分布是左偏的，此时均值左边的数据比均值右边的数据更散；当 p_d 接近 0 时，称分布无偏倚，即分布是对称的。若知道分布有可能在偏度上偏离正态分布，则可用偏度来检验分布的正态性。

数据分布右偏时，一般有算术平均数＞中位数＞众数；左偏时相反，即众数＞中位数＞算术平均数。

峰度是用来衡量数据尾部分散性的指标。

样本数据的峰度定义为：

$$f_d = \frac{n^2 u_4}{(n-1)(n-2)s^4} - \frac{3(n-1)^2}{(n-2)(n-3)} \tag{2.1.13}$$

其中，u_4、s 分别表示数据的 4 阶中心矩与标准差。

当数据的总体分布是正态分布时，峰度近似为 0；与正态分布相比较，当峰度大于 0 时，数据中含有较多远离均值的极端数值，称数据分布具有平峰厚尾性；当峰度小于 0 时，表示均值两侧的极端数值较少，称数据分布具有尖峰细尾性。在金融时间序列分析中，通常要研究数据是否具有尖峰、厚尾等特性。

在 MATLAB 中，计算样本数据偏度的命令为 skewness，调用格式为：

```
Pd= skewness(x,flg)
```

其中，输入 x 是样本数据；flg 取 0 或 1。当 flg 取 0 时，按（2.1.12）式计算偏度；当 flg 取 1 时，按公式 $\frac{1}{nS_0^3}\sum_{i=1}^{n}(X_i-\overline{X})^3$ 计算偏度，S_0 是未修正的标准差。当 x 是矩阵时，输出 Pd 为数组，其第 i 个元素是 x 的第 i 列数据的偏度。

计算样本数据的峰度的有关命令为 kurtosis，调用格式为：

```
fd= kurtosis(x,flg)- 3
```

其中，输入 x 是样本数据；flg 取 0 或 1。当 flg 取 0 时，按（2.1.13）式计算峰度；当 flg 取 1 时，按公式 $\frac{1}{nS_0^4}\sum_{i=1}^{n}(X_i-\overline{X})^4$ 计算峰度，S_0 是未修正的标准差。输出 fd 为峰度，当 x 是矩阵时，fd 为数组，其第 i 个元素是 x 的第 i 列数据的峰度。

例 2.1.4　MATLAB 系统中提供了 IBM 公司 1995 年 1 月 3 日至 1999 年 4 月 1 日的股票开盘价、最高价、最低价、收盘价和成交量数据，数据文件为“ibm9599.dat”，试计算各项数据的偏度、峰度。

解：先在 MATLAB 命令窗口中输入命令

```
ibm= ascii2fts('ibm9599.dat',1,3,2);          % 调入数据文件 ibm9599.dat
```

此时得到数据矩阵 ibm 共有 6 列，第 1 列为日期，其余各列分别为股票开盘价、最高价、最低价、收盘价和成交量数据。其他程序如下：

```
tsmat= fts2mat(ibm);                          % 提取 ibm 数据的后五列数据矩阵
pd= skewness(tsmat,0);                        % 计算 tsmat 每列数据的偏度
fd= kurtosis(tsmat,0)- 3;                     % 计算 tsmat 每列数据的峰度
[pd;fd]                                       % 输出计算结果
subplot(2,2,1),histfit(tsmat(:,1)),title('open')     % 作开盘价直方图
subplot(2,2,2),histfit(tsmat(:,2)),title('high')     % 作最高价直方图
subplot(2,2,3),histfit(tsmat(:,3)),title('low')      % 作最低价直方图
subplot(2,2,4),histfit(tsmat(:,4)),title('close')    % 作收盘价直方图
```

从表 2-5 中的数字可以看出，1995 年 1 月 3 日至 1999 年 4 月 1 日，IBM 公司股票开盘价、最高价、最低价、收盘价和成交量数据偏度均大于 0，开盘价与成交量的峰度较大，说明数据不是来自于正态分布总体。从各个指标数据的直方图（图 2-1）也可以看出，分布呈右偏态。

表 2-5　IBM 公司股票开盘价、最高（低）价、收盘价、成交量偏度与峰度

统计量	开盘价	最高价	最低价	收盘价	成交量
偏度	0.934 7	0.889 8	0.907 8	0.891 2	2.944 8
峰度	0.174 5	−0.023 6	0.001 8	−0.022 5	16.224 6

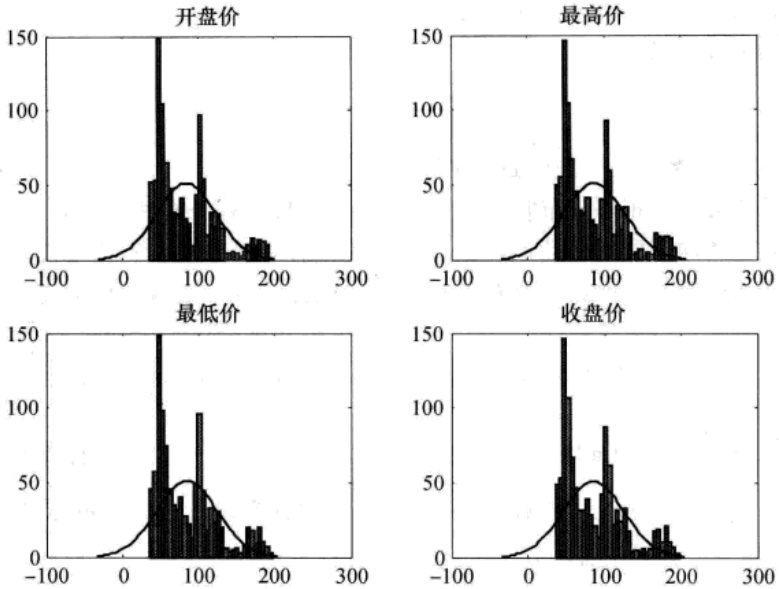


图 2-1 IBM 公司股票开盘价、最高价、最低价、收盘价直方图

2.1.2 样本数据可视化

1. 可视化

数据可视化是指数据的图形表示。借助几何图形可以形象地说明数据的特征与分布情况。常用的图形有条形图、直方图、盒图、阶梯图和火柴棒图等。

1）条形图。条形图是用宽度相同的直线条的高低或长短来表示统计指标数值的大小。条形图根据表现资料的内容可分为单式条形图、复式条形图和结构条形图。单式条形图反映统计对象随某一因素变化而变化的情况。复式条形图可以反映统计对象随两个因素变动而变动的情况。结构条形图则反映不同统计对象内部结构的变化情况。

在 MATLAB 中，绘制条形图的命令为 bar，调用格式为：

```
bar(X)                    % 作样本数据 X 的条形图
bar(x,Y)                  % x 的元素在横坐标轴上按从小到大排列,作 Y 和 x 对应的条形图
```

2）直方图。将观测数据的取值范围分为若干个区间，计算落在每个区间的频数或频率。在每个区间上画一个矩形，以估计总体的概率密度。

在 MATLAB 中，绘制直方图的命令为 hist，调用格式为：

```
hist(x,n)                 % 作数据 x 的直方图,其中 n 表示分组的个数,默认值 n= 10
```

附加有正态密度曲线的直方图命令为 histfit，调用格式为：

```
histfit(X)                % X 为样本数据向量,返回直方图和正态曲线
histfit(X,nbins)          % nbins 指定 bar 的个数,默认值为 X 中数据个数的平方根
```

3）盒图。盒图是由 5 个数值点组成：最小值、下四分位数、中位数、上四分位数、最大值。中间的盒子是从 Q1 延伸到 Q3，盒子里的直线标示出中位数的位置，盒子两端有直线往外延伸到最小数与最大数。

在 MATLAB 中，绘制盒图的命令为 boxplot，调用格式为：

```
boxplot(X)                    % 产生矩阵 X 的每一列的盒图和“须”图，“须”是从盒的尾部延伸出来，并表示盒外
                                数据长度的线，如果“须”的外面没有数据，则在“须”的底部有一个点
```

4）阶梯图。绘制阶梯图的命令为 stairs，调用格式为：

```
stairs(x)                     % 作数据 x 的阶梯图
```

5）火柴棒图。绘制火柴棒图的命令为 stem，调用格式为：

```
stem(x)                       % 作离散数据序列 x 的火柴棒图
```

例 2.1.5　随机生成 150 个服从标准正态分布的随机数，将这些数据作为样本数据，分别作出样本数据的条形图、直方图、阶梯图、火柴棒图。

解：编写程序如下：

```
x= random('normal',0,1,[1,150]);  % 产生服从标准正态分布的随机数 150 个
bar(x),legend('bar(x)')           % 作条形图(图 2-2)
hist(x,),legend('hist(x)')        % 作直方图(图 2-3)
stairs(x),legend('stairs(x)')     % 作阶梯图(图 2-4)
stem(x),legend('stem(x)')         % 作火柴棒图(图 2-5)
```

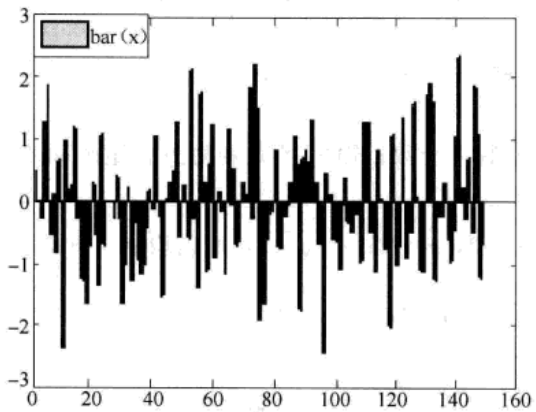


图 2-2　条形图

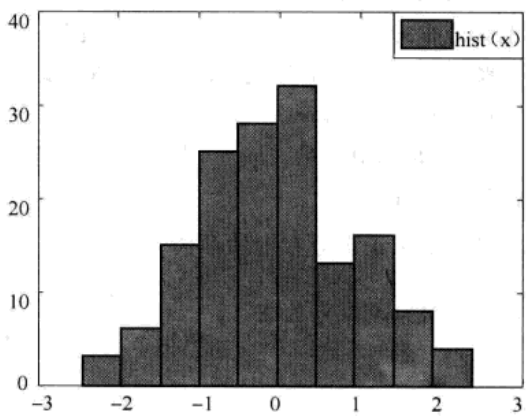


图 2-3　直方图

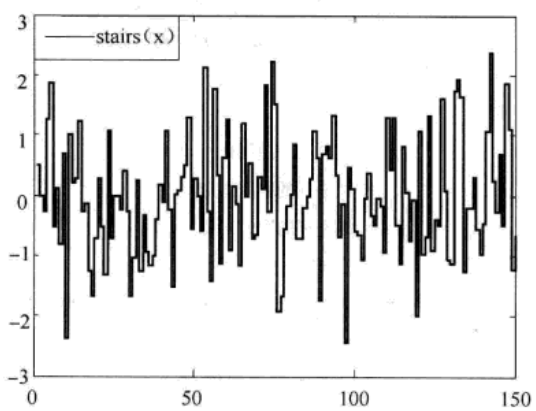


图 2-4　阶梯图

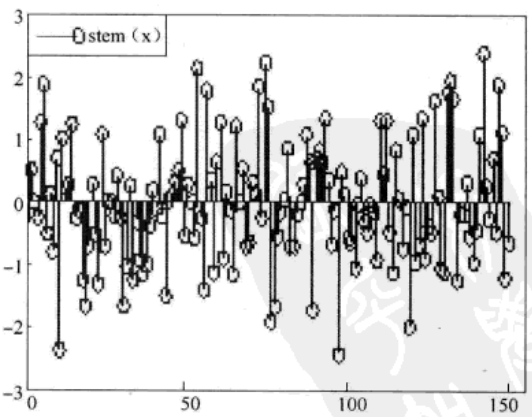


图 2-5　火柴棒图

读者还可作出盒图与附加有正态密度曲线的直方图等。

2. 二维与三维数据可视化

1）绘制散点图的命令为 scatter 与 scatter3，调用格式为：

```
scatter(x,y)
```

其中，x 是横坐标向量，y 是纵坐标向量，输出平面散点图。

```
scatter3(x,y,z)
```

其中，x、y、z 分别是横坐标、纵坐标、竖坐标向量，输出空间散点图。

2）绘制曲面图的命令为 mesh 与 surf，调用格式为：

```
mesh (X, Y, Z) 或 surf (X, Y, Z)
```

其中，Z 是对应 (X, Y) 处的函数值，即 $Z=f(X, Y)$，$[X, Y]$ 是由命令 meshgrid 生成的数据点矩阵，即 $[X, Y]$=meshgrid(x, y)，输入向量 x 为 xoy 平面上矩形定义域的矩形分割线在 x 轴上的坐标，向量 y 为 xoy 平面上矩形定义域的矩形分割线在 y 轴上的坐标。矩阵 X 为 xoy 平面上矩形定义域的矩形分割点的横坐标值矩阵，X 的每一行是向量 x，且 X 的行数等于 y 的维数；矩阵 Y 为 xoy 平面上矩形定义域的矩形分割点的纵坐标值矩阵，Y 的每一列是向量 y，且 Y 的列数等于 x 的维数。

例 2.1.6 设总体 (X, Y) 服从二维正态分布 $N(2, 1; 3, 3; \sqrt{3}/2)$，生成 100 对服从该分布随机数据对 (x_i, y_i)，将这些数据作为样本数据，绘制样本数据的散点图。再根据二维正态分布的密度函数，绘制密度曲面图。

解：随机生成服从二维正态分布的数据的命令为 mvnrnd，调用格式为：

```
X= mvnrnd(mu,sigma,n)
```

其中，mu 是均值向量；sigma 是协方差矩阵；n 是数据个数；输出 X 是和协方差矩阵同阶的随机数据矩阵。

已知二维正态分布中的参数 $\mu_1=2$，$\sigma_1^2=1$；$\mu_2=3$，$\sigma_2^2=3$；$\rho=\sqrt{3}/2$，所以均值向量为 $\mu=(2, 3)$，协方差矩阵为 $\Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}=\begin{pmatrix}1 & 1.5\\ 1.5 & 3\end{pmatrix}$，编写程序如下：

```
clear
mu= [2 3];                                          % 输入均值向量
sa= [1 1.5;1.5 3];                                  % 输入协方差矩阵
r= mvnrnd(mu,sa,100);                               % 随机生成 n= 100 的样本数据
scatter(r(:,1),r(:,2),'* ');                        % 绘制样本数据的平面散点图
% 绘制密度曲面
figure(2)
v= sqrt(3)/2;                                       % 输入相关系数
x= - 1:0.05:5;                                      % 横坐标的取值向量
y= - 2:0.05:8;                                      % 纵坐标的取值向量
[X,Y]= meshgrid(x,y);                               % 生成网格点
T= ((X- mu(1)).^2/sa(1,1)- 2*v/sqrt(sa(1,1)*sa(2,2))*(X- mu(1)).*(Y- mu(2))+ (Y- mu(2)).^2/sa
(2,2));
Z= 1/(2*pi)/sqrt(det(sa))*exp(- 1/2/(1- 3/4)*T);    % 计算密度函数值
mesh(X,Y,Z)                                         % 绘制曲面
```

输出图形结果分别如图 2-6 和图 2-7 所示。

由图 2-6 可以看出，散点图位于平面上的一个椭圆状区域内，不同的相关系数对应的椭圆状区域形状不同，相关系数越接近于 1，椭圆越扁长。我们可以利用这一图形特征初步说明数据是否来自正态总体。

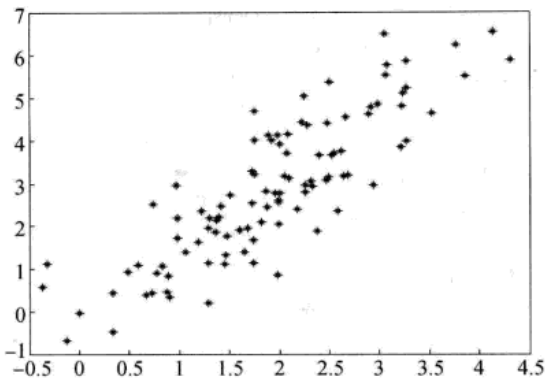


图 2-6　样本数据的散点图

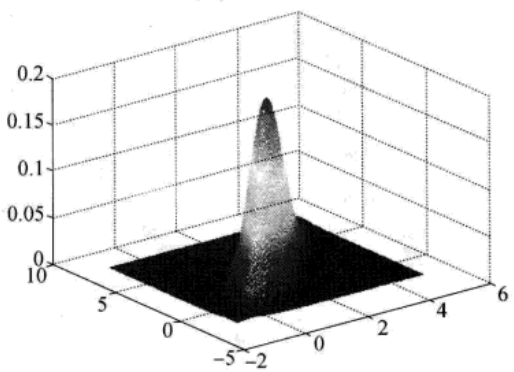


图 2-7　服从 $N(2,1;3,3;\sqrt{3}/2)$ 分布的密度曲面图

3. QQ 图

设总体服从正态分布 $N(\mu,\sigma^2)$，来自总体的样本为 $x_1,x_2,\cdots,x_n$，其次序统计量 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，则平面上 n 个点

$$\left(\Phi^{-1}\left(\frac{i-0.375}{n+0.375}\right),x_{(i)}\right)(i=1,2,\cdots,n) \tag{2.1.14}$$

的散点图称为样本 QQ 图，其中 $\Phi^{-1}(\cdot)$ 为标准正态分布函数的反函数。

可以证明，若样本确实来自正态总体，则散点在直线 $y=\sigma x+\mu$ 附近，即 QQ 图大致呈现一条直线形状。当样本来自其他分布总体时，样本 QQ 图将是弯曲的。这样，利用 QQ 图可以直观地作正态性检验，即若 QQ 图近似一条直线，则可认为样本数据来自正态总体。

对于其他类型的分布，也有相应的 QQ 图，其中散点的横坐标为该分布的对应分位数，以此可判断数据是否近似服从该类型的分布。

在 MATLAB 中，作正态分布 QQ 图的命令为 normplot，调用格式为：

```
normplot(X)
```

其中，当输入 X 为向量时，显示正态分布 QQ 图；当 X 为矩阵时，显示每一列的正态分布概率图。

作威布尔（Weibull）分布的 QQ 图的命令为 weibplot，调用格式为：

```
weibplot(X)
```

其中，若输入 X 为向量，则显示威布尔分布 QQ 图；若 X 为矩阵，则显示每一列的威布尔概率图。

如果数据点基本散布在直线上，则表明数据服从该分布，否则拒绝该分布。

例 2.1.7　对于例 2.1.6 模拟的样本数据 r，分别作出两个分量的 QQ 图，从 QQ 图检验各分量是否服从正态分布。

解： 编写程序如下：

```
subplot(1,2,1),normplot(r(:,1))        % 分量 x 的 QQ 图
subplot(1,2,2),normplot(r(:,2))        % 分量 y 的 QQ 图
```

从图 2-8 可以看出，两个分量的 QQ 图呈现一条直线形状，所以样本中的每个分量都服从正态分布。事实上，数学理论上已证明：二元正态分布的边际分布仍为正态分布。QQ 图反映的结果与理论相一致。

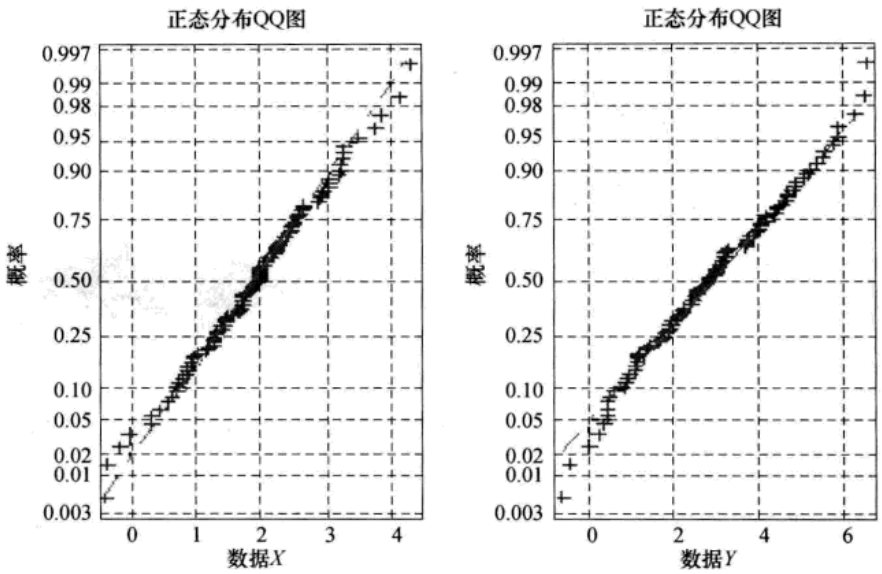


图 2-8 两个分量的正态分布 QQ 图

2.2 数据分布及检验

样本数据的数字特征刻画了数据的主要特征，而要对数据的总体情况作全面地了解，就必须研究数据的分布。上一节中的数据直方图与 QQ 图等能直观粗略描述数据的分布，本节进一步研究如何判定数据是否服从正态分布的问题。如果不服从正态分布，那么又可能服从怎样的分布呢？

2.2.1 一元数据分布检验

1. 经验分布函数

设来自总体 X 的容量为 n 的样本 $x_1, x_2, \cdots, x_n$，样本的次序统计量为 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$，对于任意实数 x，定义函数

$$F_n(x) = \begin{cases} 0 & \text{若 } x < x_{(1)} \\ \dfrac{k}{n} & \text{若 } x_{(k)} \leqslant x < x_{(k+1)} \\ 1 & \text{若 } x \geqslant x_{(n)} \end{cases} \tag{2.2.1}$$

称 $F_n(x)$ 为经验分布函数。

由定义，$F_n(x)$ 表示事件 $\{X \leqslant x\}$ 在 n 次独立重复试验中的频率。

1933 年，格利文科（Glivenko）证明了以下的结果：对于任一实数 x，当 $n \to \infty$ 时，$F_n(x)$ 以概率 1 一致收敛于分布函数 $F(x)$，即

$$P\{\lim_{n\to\infty} \sup_{-\infty<x<\infty} |F_n(x) - F(x)| = 0\} = 1$$

这一结论表明：对于任一实数 x，当 n 充分大时，

$$F(x) \approx F_n(x) \tag{2.2.2}$$

因此，可用经验分布函数来近似代替 $F(x)$，这也是通过样本推断总体的最基本理论依据

之一。

在 MATLAB 中，作经验（累积）分布函数图的命令为 cdfplot，调用格式为：

```
cdfplot(X)                          % 作样本 X(向量)的经验(累积)分布函数图
h= cdfplot(X)                       % h 表示曲线的环柄
[h,stats]= cdfplot(X)               % 输出 stats 表示样本最小值、最大值、均值、
                                      中值与标准差
```

通常，将样本的直方图与经验分布函数图结合应用，对数据的分布作出推断。

例 2.2.1　生成服从标准正态分布的 50 个样本点，作出样本的经验分布函数图，并与理论分布函数 $\Phi(x)$ 比较。

解： 编写程序如下：

```
X= normrnd(0,1,50,1);               % 生成服从标准正态分布的 50 个样本点
[h,stats]= cdfplot(X);              % 作样本的经验分布函数图
hold on
plot(- 3:0.01:3,normcdf(- 3:0.01:3,0,1),'r')   % 作理论分布函数图
```

输出结果：

```
h=
   3.0013
stats=
       min:- 1.8740                 % 样本最小值
       max:1.6924                   % 最大值
       mean:0.0565                  % 均值
       median:0.1032                % 中值
       std:0.7559                   % 样本标准差
```

图 2-9 表明经验分布函数图形与理论分布函数图很相近。

2. 总体分布的正态性检验

进行参数估计和假设检验时，通常总是假定总体服从正态分布。虽然在许多情况下这个假定是合理的，但是当要以此为前提进行重要的参数估计或假设检验，或者人们对它有较大怀疑的时候，就有必要对这个假设进行检验。进行总体正态性检验的方法有很多种，下面针对 MATLAB 统计工具箱中提供的程序，简单介绍几种方法。

（1）Jarque-Bera 检验

Jarque-Bera 检验简称 JB 检验，它是利用正态分布的偏度 g_1 和峰度 g_2，构造一个包含 g_1、g_2 且自由度为 2 的卡方分布统计量 JB，即

$$JB = n\left(\frac{1}{6}J^2 + \frac{1}{24}B^2\right) \sim \chi^2(2) \quad (2.2.3)$$

其中 $J = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3$，$B = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4 - 3$。

对于显著性水平 α，当 JB 统计量小于 χ^2 分布的 $1-\alpha$ 分位数 $\chi^2_{1-\alpha}(2)$ 时，接受 H_0，即认为总体服从正态分布；否则拒绝 H_0，即认为总体不服从正态分布。这个检验适用于大样本，当样本容量 n 较小时需慎用。

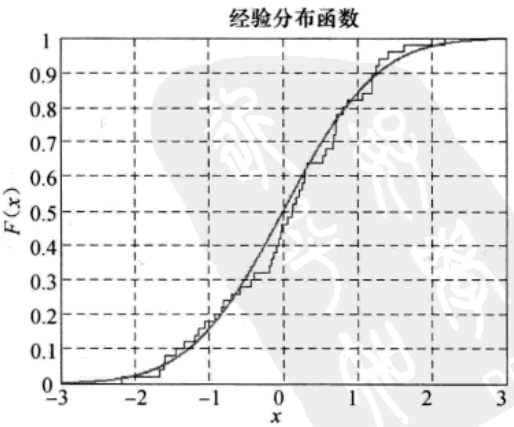


图 2-9　$N(0, 1)$ 分布函数图及其 50 个样本点的经验分布函数图

在 MATLAB 中，JB 检验的命令为 jbtest，调用格式为：

```
H= jbtest(X,alpha)或[H,P,JBSTAT,CV]= jbtest(X,alpha)
```

对输入向量 X 进行 Jarque-Bera 检验，显著性水平 α 默认值为 0.05。P 为接受假设的概率值，JBSTAT 为测试统计量的值，CV 为是否拒绝原假设的临界值。H 为测试结果，若 $H=0$，则认为 X 服从正态分布；若 $H=1$，则可以否定 X 服从正态分布。P 小于 α，则拒绝是正态分布的原假设；JBSTAT 大于 CV 可以拒绝是正态分布的原假设；命令 jbtest 一般用于大样本，对于小样本用命令 lillietest。

(2) Kolmogorov-Smirnov 检验

Kolmogorov-Smirnov 检验简称 KS 检验，它是通过样本的经验分布函数与给定分布函数的比较，推断该样本是否来自给定分布函数的总体。设给定分布函数为 $G(x)$，构造统计量

$$D_n = \max_n(|F_n(x) - G(x)|) \tag{2.2.4}$$

即两个分布函数之差的最大值，对于假设 H_0：总体服从给定的分布 $G(x)$ 和 α，根据 D_n 的极限分布确定统计量关于是否接受 H_0 的临界值。

因为这个检验需要给定 $G(x)$，所以当用于正态性检验时，只能做标准正态检验，即 H_0：总体服从标准正态分布 $N(0, 1)$。

在 MATLAB 中，KS 检验命令为 kstest，调用格式为：

```
h= kstest(x)
h= kstest(x,cdf)
[h,p,ksstat,cv]= kstest(x,cdf,alpha)
```

把向量 x 中的值与标准正态分布进行比较并返回假设检验结果 h。如果 $h=0$，则表示不能拒绝原假设，即不能拒绝服从正态分布。假设的显著水平默认值是 0.05。cdf 是一个两列矩阵，矩阵的第一列包含可能的 x 值，第二列是假设累积分布函数 $G(x)$ 的值，在可能的情况下，cdf 的第一列应包含 x 中的值，如果第一列没有，则用插值的方法近似。指定显著水平 alpha，返回 p 值，KS 检验统计量 Ksstat；临界值 cv。

(3) Lilliefors 检验

Lilliefors 检验是改进的 KS 检验并用于一般的正态性检验，原假设 H_0：总体服从正态分布 $N(\mu, \sigma^2)$，其中 μ、σ^2 由样本均值和方差估计。

该检验的 MATLAB 命令为 lillietest，调用格式为：

```
H= lillietest(X,alpha)或[H,P,LSTAT,CV]= lillietest(X,alpha)
```

对输入向量 X 进行 Lilliefors 测试，显著性水平 alpha 在 0.01～0.2 之间，默认值为 0.05。输出 P 为接受假设的概率值，LSTAT 为测试统计量的值，CV 为是否拒绝原假设的临界值。H 为测试结果，若 $H=0$，则 X 服从正态分布；若 $H=1$，则可以否定 X 服从正态分布。若 P 小于 alpha，则可以拒绝是正态分布的原假设；JBSTAT 大于 CV 可以拒绝是正态分布的原假设。

2.2.2 多维数据的特征值与分布检验

1. 多维数据的数字特征

设总体为 p 维向量 $G=(X_1, X_2, \cdots, X_p)$，从中抽取样本容量为 n 的样本，第 i 个样本观测值为 $x_i=(x_{i1}, x_{i2}, \cdots, x_{ip})$ $(i=1, 2, \cdots, n)$，记

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{2.2.5}$$

称 X 为样本数据矩阵。为了方便起见，X 的第 j 个列向量记为 $X_j=(x_{1j},\ x_{2j},\ \cdots,\ x_{nj})^{\mathrm{T}}$。

显然，X 的第 j 个列向量是 X_j 的 n 个观测数据。通常由样本数据矩阵 X 出发，构造下列统计量来分析总体的特征。

1）样本均值向量。记 X_j 的观测值（即 X 中的第 j 列）的均值为

$$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}\quad(j=1,2,\cdots,p) \tag{2.2.6}$$

则 $\overline{x}=(\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_p)^{\mathrm{T}}$ 称为 p 元样本均值向量。

2）样本协方差矩阵。记

$$s_{jk}=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)(x_{ik}-\overline{x}_k)\quad(j,k=1,2,\cdots,p) \tag{2.2.7}$$

则 s_{jk} 称为 X_j 与 X_k 的样本协方差，或称为样本数据矩阵 X 的第 j 列与第 k 列的协方差。记

$$S=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \tag{2.2.8}$$

则 S 称为样本协方差矩阵。

显然，X_j 与 X_j 的协方差为 s_{jj}，即

$$s_{jj}=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2\quad(j=1,2,\cdots,p) \tag{2.2.9}$$

3）样本相关系数矩阵。X 的第 j 列与第 k 列的相关系数记为

$$r_{jk}=\frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}\quad(j,k=1,2,\cdots,p)$$

又记

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix}=\begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{2.2.10}$$

则 R 称为样本相关系数矩阵。

不难验证，样本相关系数矩阵与样本协方差矩阵存在如下关系：

$$R=C^{\mathrm{T}}SC \tag{2.2.11}$$

其中 $C=\begin{pmatrix} \sqrt{s_{11}}^{-1} & & & \\ & \sqrt{s_{22}}^{-1} & & \\ & & \ddots & \\ & & & \sqrt{s_{pp}}^{-1} \end{pmatrix}$。

4）样本标准化矩阵。令

$$x_{ij}^{*}=\frac{x_{ij}-\overline{x}_j}{\sqrt{s_{jj}}}(i=1,2,\cdots,n;j=1,2,\cdots,p) \tag{2.2.12}$$

则

$$X^{*}=(x_{ij}^{*})_{n\times p} \tag{2.2.13}$$

称为样本矩阵 X 的标准化矩阵。

可以证明：X^{*} 的协方差矩阵 S^{*} 等于 X 的相关系数矩阵 R，即 $S^{*}=R$。

5）R 矩阵。X 的第 j 列与第 k 列的 R 系数定义为

$$r_{jk}=\frac{2[X_j,X_k]}{[X_j,X_j]+[X_k,X_k]} \tag{2.2.14}$$

其中 $[X_j,X_k]=\sum_{i=1}^{n}x_{ij}x_{ik}(j,k=1,2,\cdots,p)$，则矩阵 $(r_{jk})_{p\times p}$ 称为矩阵 X 的 R 矩阵，记为 $R(X)$，即

$$R(X)=(r_{jk})_{p\times p}$$

由式（2.2.14），显然 $r_{jj}=1$（$j=1,2,\cdots,p$），$|r_{jk}|\leqslant 1$（$j,k=1,2,\cdots,p$）。

可以证明：对于矩阵 X，有 $R(X^{*})=R$，即 X 的标准化矩阵的 R 矩阵等于其相关系数矩阵。

协方差矩阵与量纲有关，相关系数矩阵和 R 矩阵与量纲无关，这一点在今后的判别分析中值得注意。

协方差矩阵、相关系数矩阵与 R 矩阵都是实对称非负定矩阵。

在 MATLAB 中，计算样本协方差矩阵的命令为 cov，调用格式为：

```
S= cov(X)
```

当 X 为向量时，S 表示 X 的方差；当 X 为矩阵时，S 表示 X 的协方差矩阵，即 S 的对角线元素是 X 每列的方差，S 的第 i 行第 j 列元素为 X 的第 i 列和第 j 列的协方差值。

计算样本相关系数矩阵的命令为 corrcoef，调用格式为：

```
R= corrcoef(X)
```

其中 X 为样本矩阵，输出 R 的对角线元素是 1，R 的第 i 行第 j 列元素为 X 的第 i 列和第 j 列的相关系数。

计算 X 的标准化矩阵命令为 zscore，调用格式为：

```
Z= zscore(X)
```

其中 X 为样本矩阵，输出 Z 是标准化矩阵。

MATLAB 中没有计算 R 矩阵的命令，因此根据 R 矩阵的定义，可编写计算 R 矩阵的程序如下：

```
% 输入样本数据矩阵 X
X= [data];
% 计算 R 矩阵
for i= 1:size(X,2)
        for j= 1:size(X,2)
          RX(i,j)= 2 * dot(X(:,i),X(:,j))./[sum(X(:,i).^2)+ sum(X(:,j).^2)];
        end
end
RX                  % 输出 R(X)
```

2. 多维正态分布的概念与性质

设 p 维总体 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的分布密度函数为

$$f(x_1,x_2,\cdots,x_p)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right\} \tag{2.2.15}$$

则称 X 服从 p 维正态分布，记为 $X\sim N_p(\mu,\Sigma)$，其中

$$x=(x_1,x_2,\cdots x_p)^{\mathrm{T}},\mu=(\mu_1,\mu_2,\cdots\mu_p)^{\mathrm{T}},\Sigma=\begin{pmatrix}\sigma_{11}&\sigma_{12}&\cdots&\sigma_{1p}\\\sigma_{21}&\sigma_{22}&\cdots&\sigma_{2p}\\\cdots&\cdots&\cdots&\cdots\\\sigma_{p1}&\sigma_{p2}&\cdots&\sigma_{pp}\end{pmatrix}$$

μ 称为总体均值向量，Σ 称为总体协方差矩阵。

多维正态分布具有以下性质：

1）多维正态分布的边缘分布服从正态分布，但反之不真。

2）正态随机向量的线性函数仍然服从正态分布，即若 $X\sim N_p(\mu,\Sigma)$，A 为 $s\times p$ 阶常数矩阵，d 为 s 维常数向量，则

$$AX+d\sim N_s(A\mu+d,A\Sigma A^{\mathrm{T}})$$

即多维正态分布在线性变换下仍然服从多维正态分布。

3）正态分布的随机向量间相互独立与不相关等价。

对于来自总体且由（2.2.5）式表示的样本数据矩阵 X，怎样检验其是否是来自于多维正态总体呢？一般可按照以下 QQ 图检验方法，具体的过程如下：

1）由样本数据矩阵 X 计算均值向量 $\overline{X}$ 和协方差矩阵 S。

2）计算顺序统计量 $X_{(t)}$ 到 $\overline{X}$ 的马氏平方距离（参见第 4 章第 4.1 节）

$$D_t^2=(X_{(t)}-\overline{X})^{\mathrm{T}}\Sigma^{-1}(X_{(t)}-\overline{X})\quad(t=1,\cdots,n)$$

3）对上述马氏平方距离从小到大排序

$$D_{(1)}^2\leqslant D_{(2)}^2\leqslant\cdots\leqslant D_{(n)}^2$$

4）计算 $p_t=\dfrac{t-0.5}{n}(t=1,2,\cdots,n)$ 及 χ_t^2，其中 χ_t^2 满足 $H(\chi_t^2\mid p)=p_t$。

5）以马氏距离为横坐标，χ_t^2 分位数为纵坐标作 n 个点 $(D_{(t)}^2,\chi_t^2)$ 的平面散点图，即分布的 QQ 图。

6）考察散点图是否在一条通过原点且斜率为 1 的直线上，若是，则接受数据来自 p 元正态分布总体的假设，否则拒绝正态分布假设。

以上 QQ 图检验方法的 MATLAB 程序实现如下：

```
% 输入样本数据矩阵 X
X= [data];
[N,p]= size(X);                          % X 的行数及列数
d= mahal(X,X);                           % 计算马氏距离
d1= sort(d);                             % 马氏距离从小到大排序
pt= [[1:N]- 0.5]/N;                      % 计算分位数
x2= chi2inv(pt,p);                       % 计算 χ²t
plot(d1,x2','* ',[0:m],[0:m],'- r')      % 作散点图与直线 y= x,其中 m 是正整数
```

以下举例说明上述程序的应用。

例 2.2.2 为了研究某种疾病，对一批 60 人分为 3 组：G_1、G_2、G_3，同时进行 4 项指标的检测：β脂蛋白（X_1），甘油三酯（X_2），α脂蛋白（X_3），前β脂蛋白（X_4），检测的结果列在表 2-6 中，现将 3 组检验数据视为一个总体，问总体是否服从四维正态分布？

表 2-6 4 项指标检测数据

*G*1				*G*2				*G*3			
X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
260	75	40	18	310	122	30	21	320	64	39	17
200	72	34	17	310	60	35	18	260	59	37	11
240	87	45	18	190	40	27	15	360	88	28	26
170	65	39	17	225	65	34	16	295	100	36	12
270	110	39	24	170	65	37	16	270	65	32	21
205	130	34	23	210	82	31	17	380	114	36	21
190	69	27	15	280	67	37	18	240	55	42	10
200	46	45	15	210	38	36	17	260	55	34	20
250	117	21	20	280	65	30	23	260	110	29	20
225	130	36	11	200	76	39	20	240	114	38	18
210	125	26	17	280	94	26	11	310	103	32	18
170	64	31	14	190	60	33	17	330	112	21	11
270	76	33	13	295	55	30	16	345	127	24	20
190	60	34	16	270	125	24	21	250	62	22	16
280	81	20	18	280	120	32	18	260	59	21	19
310	119	25	15	240	62	32	20	225	100	34	30
270	57	31	8	280	69	29	20	345	120	36	18
250	67	31	14	370	70	30	20	360	107	25	23
260	135	39	29	280	40	37	17	250	117	36	16

数据来源：高惠璇．应用多元统计分析［M］．北京：北京大学出版社，2005.

解：先将表 2-6 中数据按原位置作为矩阵 A 输入，然后整理成样本数据矩阵 X，程序如下：

```
A=[260  75  40  18  310  122  30  21  320  64  39  17;
   200  72  34  17  310   60  35  18  260  59  37  11;
   ...
   260 135  39  29  280   40  37  17  250 117  36  16];
X=[A(:,1:4);A(:,5:8);A(:,9:12)];
[N,p]=size(X);
d=mahal(X,X);                          % 计算马氏距离
d1=sort(d);                            % 从小到大排序
pt=[[1:N]-0.5]/N;                      % 计算分位数
x2=chi2inv(pt,p);                      % 计算 χ²
plot(d1,x2','*',[0:12],[0:12],'-r')    % 作图
```

输出图形如图 2-10 所示。

从图 2-10 可以看出，数据点基本落在直线上，故不能拒绝该数据服从四维正态分布的假设。

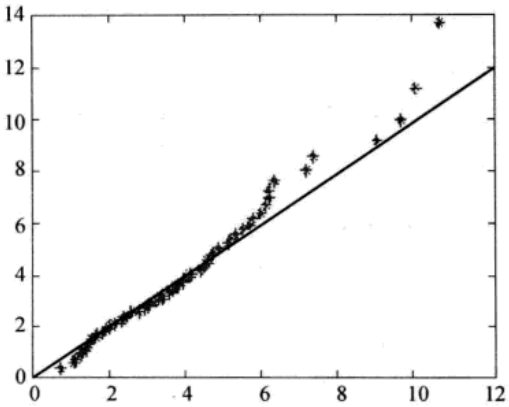


图 2-10　4 项检测数据的多维正态检验 QQ 图

3. 多维数据的多个总体协方差矩阵的相等性检验

(1) 两个总体协方差矩阵相等的检验

设从两个总体分别抽取样本容量为 n_1、n_2 的两个样本，其样本的协方差矩阵分别为 S_1、S_2，那么在两个总体协方差矩阵相等时，其总体的协方差矩阵的估计为：

$$S=\frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$$

若检验两个总体的协方差矩阵相等，可以有如下的假设检验：

$$H_0:S_i=S\leftrightarrow H_1:S_i\neq S,(i=1,2)$$

检验统计量

$$Q_i=(n_i-1)[\ln|S|-\ln|S_i|-p+\mathrm{tr}(S^{-1}S_i)]\sim\chi^2(p(p+1)/2)\quad(i=1,2)\qquad(2.2.16)$$

其中 $|\cdot|$ 表示行列式，p 是向量的维数，tr 表示矩阵的迹。

对给定的 α，卡方分布临界值为 λ，若 $Q_i<\lambda$，($i=1$，2) 则接受 H_0，否则拒绝 H_0。

(2) 多个总体协方差矩阵相等的检验

设有 k 个 p 元总体 $G_i(i=1,2,\cdots,k)$，从中抽取样本容量为 $n_i(i=1,2,\cdots,k)$ 的 k 个样本，其样本的协方差矩阵为 $S_1,S_2,\cdots,S_k$，用 $S_1,S_2,\cdots,S_k$ 估计 $\Sigma_1,\Sigma_2,\cdots,\Sigma_k$。

原假设 $H_0:\Sigma_1=\Sigma_2=\cdots=\Sigma_k$；

备择假设 $H_1:\Sigma_1,\Sigma_2,\cdots,\Sigma_k$ 至少有一对不相等。

在 H_0 成立时，统计量

$$\xi=(1-d)M\sim\chi^2(f)\qquad(2.2.17)$$

其中

$$d=\begin{cases}\dfrac{2p^2+3p-1}{6(p+1)(k-1)}\left[\displaystyle\sum_{i=1}^{k}\frac{1}{n_i-1}-\frac{1}{n-k}\right] & n_i\text{ 不全等}\\ \dfrac{(2p^2+3p-1)(k+1)}{6(p+1)(n-k)} & n_i\text{ 全等}\end{cases}\qquad(2.2.18)$$

$M=(n-k)\ln|S|-\sum_{i=1}^{k}(n_i-1)\ln|S_i|$，$S=\sum_{i=1}^{k}(n_i-1)S_i/(n-k)$，$f=p(p+1)(k-1)/2$ 为自由度，$n=n_1+n_2+\cdots+n_k$。

对给定的 α，计算概率 $p=P(\xi>\chi_\alpha^2(f))$，若 $p<\alpha$ 则拒绝 H_0，否则接受 H_0。

以上过程程序可用下例说明。

例 2.2.3　检验表 2-6 中 3 个总体 G_1、G_2、G_3 的协方差矩阵是否相等 ($\alpha=0.1$)？

解：编写程序如下：

```
% 输入数据
A= [data];
G1= A(:,1:4);
G2= A(:,5:8);
G3= A(:,9:12);
```

```
n= size(G1,1)+size(G2,1)+size(G2,1);                    % 计算总的样本容量
[n1,p]= size(G1);
k= 3;
f= p* (p+1) * (k-1)/2;                                  % 统计量自由度
d= (2* p^2+3* p-1) * (k+ 1)/(6* (p+1) * (n-k));         % 由(2.2.18)式计算
s1= cov(G1);                                            % 协方差矩阵
s2= cov(G2);                                            % 协方差矩阵
s3= cov(G3);                                            % 协方差矩阵
s= (n1-1) * (s1+s2+s3)/(n- k);                          % 总体协方差估计
M= (n-k) * log(det(s))-19* (log(det(s1))+log(det(s2))+log(det(s3)));
T= (1-d) * M;                                           % 统计量(2.2.17)
P0= 1-chi2cdf(T,f);                                     % 卡方分布概率
```

输出结果：

```
T=20.3316,P0=0.4374
```

由于由统计量计算得到的概率为 $P0=0.4374>0.1$，故判定 3 个总体协方差矩阵相等。

2.3 数据变换

2.3.1 数据属性变换

在解决经济问题综合评价时，评价指标通常分为效益型、成本型、适度型等类型，效益型指标值越大越好、成本型指标值越小越好、适度型指标值既不太大也不太小为好。

一般说来，对问题进行综合评价，必须统一评价指标的属性，进行指标的无量纲化处理。常见的处理方法有极差变换、线性比例变换、样本标准化变换等方法。

我们将（2.2.5）式表示的样本数据矩阵 X 的每一列理解为评价指标，共有 p 项指标，X 的每一行理解为不同决策方案关于 p 项评价指标的指标值，共有 n 个方案，这样表示第 i 个方案关于第 j 项评价指标的指标值为 $x_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,p)$。

1. 统一趋势化与无量纲化

我们用 I_1、I_2、I_3 分别表示效益型、成本型和适度型指标集合，运用极差变换法建立无量纲的优属度效益型矩阵 B 与成本型矩阵 C，运用线性比例变换法可建立无量纲的优属度效益型矩阵 D 与优属度成本型矩阵 E。

（1）效益型矩阵

其变换公式为

$$B=(b_{ij})_{n\times p},b_{ij}=\begin{cases}\dfrac{(x_{ij}-\min\limits_j x_{ij})}{(\max\limits_j x_{ij}-\min\limits_j x_{ij})} & x_{ij}\in I_1\\ \dfrac{(\max\limits_j x_{ij}-x_{ij})}{(\max\limits_j x_{ij}-\min\limits_j x_{ij})} & x_{ij}\in I_2\\ \dfrac{(\max\limits_j|x_{ij}-\alpha_j|-|x_{ij}-\alpha_j|)}{\max|x_{ij}-\alpha_j|-\min\limits_j|x_{ij}-\alpha_j|} & x_{ij}\in I_3\end{cases}\tag{2.3.1}$$

其中 α_j 为第 j 项指标的适度数值。

显然指标经过极差变换后，均有 $0 \leqslant b_{ij} \leqslant 1$，且各指标下最好结果的属性值 $b_{ij}=1$，最坏结果的属性值 $b_{ij}=0$。指标变换前后的属性值成比例。

(2) 成本型矩阵

其变换公式为

$$C=(c_{ij})_{n\times p}, c_{ij}=\begin{cases} \dfrac{(\max\limits_j x_{ij}-x_{ij})}{(\max\limits_j x_{ij}-\min\limits_j x_{ij})} & x_{ij}\in I_1 \\ \dfrac{(x_{ij}-\min\limits_j x_{ij})}{(\max\limits_j x_{ij}-\min\limits_j x_{ij})} & x_{ij}\in I_2 \\ \dfrac{|x_{ij}-\alpha_j|-\min\limits_j|x_{ij}-\alpha_j|}{\max\limits_j|x_{ij}-\alpha_j|-\min\limits_j|x_{ij}-\alpha_j|} & x_{ij}\in I_3 \end{cases} \tag{2.3.2}$$

其中 α_j 为第 j 项指标的适度数值。

显然指标经过极差变换后，均有 $0 \leqslant c_{ij} \leqslant 1$，且各指标下最坏结果的属性值 $c_{ij}=1$，最好结果的属性值 $c_{ij}=0$。

(3) 优属度效益型矩阵

其变换公式为

$$D=(d_{ij})_{n\times p}, d_{ij}=\begin{cases} \dfrac{x_{ij}}{\max\limits_j x_{ij}} & x_{ij}\in I_1 \\ \dfrac{\min\limits_j x_{ij}}{x_{ij}} & x_{ij}\in I_2 \\ \dfrac{\min\limits_j|x_{ij}-\alpha_j|}{|x_{ij}-\alpha_j|} & x_{ij}\in I_3 \end{cases} \tag{2.3.3}$$

其中 α_j 为第 j 项指标的适度数值。

(4) 优属度成本型矩阵

其变换公式为

$$E=(e_{ij})_{n\times p}, e_{ij}=\begin{cases} \dfrac{\min\limits_j x_{ij}}{x_{ij}} & x_{ij}\in I_1 \\ \dfrac{x_{ij}}{\max\limits_j x_{ij}} & x_{ij}\in I_2 \\ \dfrac{|x_{ij}-\alpha_j|}{\max\limits_j|x_{ij}-\alpha_j|} & x_{ij}\in I_3 \end{cases} \tag{2.3.4}$$

其中 α_j 为第 j 项指标的适度数值。显然指标变换前后的属性值成比例。

2. 压缩变换模糊化

对于实际数据还可以通过如下的变换将原始数据压缩到 [0，1] 区间，从而构造出模糊集合。

利用 MATLAB 软件中的模糊数学工具箱，可以直接调用表 2-7 中的函数实现数据转换。

表 2-7 模糊工具箱隶属度函数

函数名称	函数表达式	命令格式	数据类型
高斯型函数	$y=e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	y=gaussmf(x, [sig, c])	适度型
钟型函数	$y=\frac{1}{1+\left\|\frac{x-c}{a}\right\|^{2b}}$	y=gbellmf(x, [a, b, c])	
S型函数	$f(x,a,b)=\begin{cases}0 & x\leqslant a\\ 2\left(\frac{x-a}{b-a}\right)^2 & a\leqslant x\leqslant\frac{a+b}{2}\\ 1-2\left(\frac{b-x}{b-a}\right)^2 & \frac{a+b}{2}\leqslant x\leqslant b\\ 1 & x\geqslant b\end{cases}$	y=smf (x, [a, b])	效益型
Z型函数	$f(x,a,b)=\begin{cases}1 & x\leqslant a\\ 1-2\left(\frac{x-a}{b-a}\right)^2 & a\leqslant x\leqslant\frac{a+b}{2}\\ 2\left(\frac{b-x}{b-a}\right)^2 & \frac{a+b}{2}\leqslant x\leqslant b\\ 0 & x\geqslant b\end{cases}$	y=zmf (x, [a, b])	成本型
sigmoid 函数	$y=\frac{1}{1+e^{-a(x-c)}}$	y=sigmf (x, [a, c])	a>0 效益 a<0 成本

2.3.2 box-cox 变换

当时间序列数据在左（或右）边有长尾巴或很不对称时，有时需要对数据进行变换以符合非参数（或参数）统计推断方法的某些条件。其中最常用的一种方法就是 box-cox 变换

$$y=\begin{cases}(x^{\lambda}-1)/\lambda & \lambda\neq 0\\ \log(x) & \lambda=0\end{cases} \tag{2.3.5}$$

在 MATLAB 中，box-cox 变换的命令为 boxcox，调用格式如下：

```
[transdat,lambda]= boxcox(x)
```

其中 x 是原始数据，transdat 是变换以后的数据，lambda 是变换公式中参数 λ 的数值。

例 2.3.1 淮河流域包括河南、安徽、江苏、山东 4 省份，1952—1991 年因水灾造成的流域成灾面积数据见表 2-8，应用 box-cox 变换考察数据的正态分布特性。

表 2-8 淮河流域成灾面积 （单位：10^6 hm^2）

年份	1952	1953	1954	1955	1956	1957	1958	1959
成灾面积	1.496 3	1.341 1	4.082	1.278 7	4.154 9	3.635 9	0.941 6	0.208 3
年份	1960	1961	1962	1963	1964	1965	1966	1967
成灾面积	1.456 7	0.856 9	2.719 7	6.749 4	3.688 4	2.539 5	0.259 6	0.274 7
年份	1968	1969	1970	1971	1972	1973	1974	1975
成灾面积	0.539 8	0.580 4	0.703 8	0.967 9	1.021 9	0.510 6	1.325 3	1.843 8
年份	1976	1977	1978	1979	1980	1981	1982	1983
成灾面积	0.493 3	0.343 7	0.285 6	2.529 6	1.659 4	0.161 5	3.208	1.469 8
年份	1984	1985	1986	1987	1988	1989	1990	1991
成灾面积	2.938	1.923 3	0.749 8	0.793 3	0.127 6	1.485 3	1.386	4.622 6

数据来源：自然灾害学报，2005，6.

解：考察正态分布特性，可检验数据是否服从正态分布或考察经验分布函数与正态分布函数的差异。将淮河流域 1952—1991 年的成灾面积数据作为矩阵 X 输入，程序如下：

```
% 绘制 QQ 图
X= [data];                          % 输入原始成灾面积数据 data
[b,t]= boxcox(X');                  % 对数据作 boxcox 变换
normplot(X)                         % 原始数据 QQ 图
figure(2)
normplot(b)                         % 变换数据 QQ 图
% 变换前后数据的经验分布函数图及相应的统计量
sa= sort(X);                        % 原始数据次序统计量
sb= sort(b);                        % 变换数据次序统计量
figure(3)
cdfplot(X);                         % 原始数据经验分布
hold on;
plot(sa,normcdf(sa),'- r')          % 正态分布函数
figure(4)
cdfplot(b);                         % 变换数据经验分布
hold on;
plot(sb,normcdf(sb),'- r')          % 变换数据经验分布与正态分布函数
```

作出图形如图 2-11 和图 2-12 所示，可以看出原始数据与正态分布函数相差甚远，变换后的数据则比较接近。

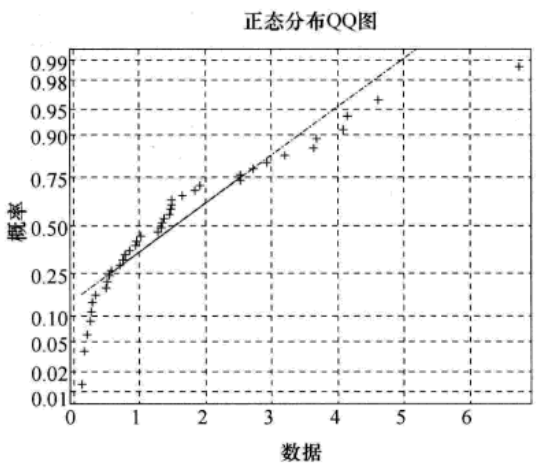


图 2-11　淮河流域成灾面积原始数据 QQ 图

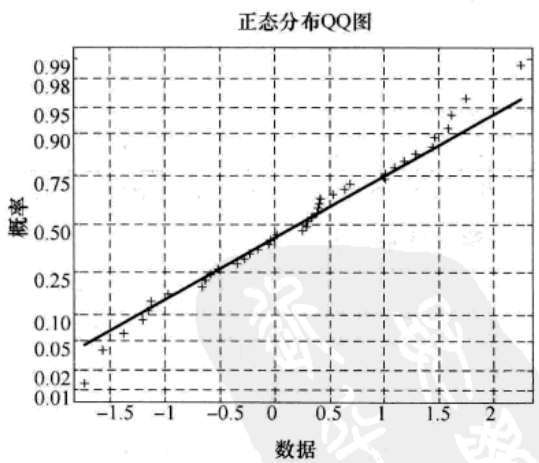


图 2-12　淮河流域成灾面积 box-cox 变换数据 QQ 图

从图形可以看出原始数据（图 2-11）没有分布在直线上，而变换后的数据（图 2-12）基本落在直线上，因此可认为原始数据不服从正态分布，而变换后的数据服从正态分布。

从图 2-13 和图 2-14 的经验分布图可以看出原始数据不服从正态分布，而变换数据近似服从正态分布。

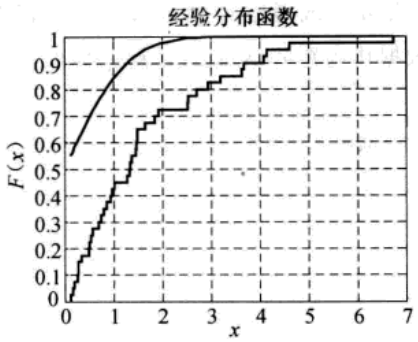


图 2-13 原始数据经验分布图

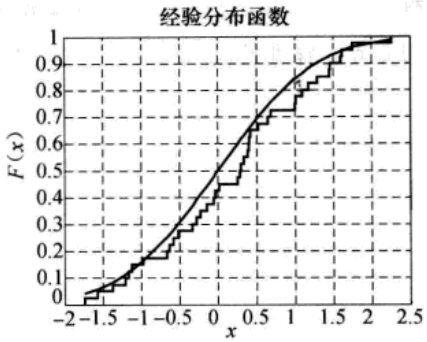


图 2-14 变换数据经验分布图

2.3.3 基于数据变换的综合评价模型

例 2.3.2 为了全面了解 10 家上市公司的绩效，用 X_1 表示每股净收益；X_2 表示净资产收益率；X_3 表示主营业务收益率；X_4 表示主营业务增长率；X_5 表示净资产增长率；X_6 表示总资产增长率。这些指标的统计数据见表 2-9，试对上市公司进行综合评价。

表 2-9 10 家上市公司的统计数据

公司编号	X_1	X_2	X_3	X_4	X_5	X_6
1	0.021	26.806	57.311	−39.815	−39.815	8.819
2	−0.142	−7.179	16.335	−11.359	−4.766	−4.626
3	−0.737	−62.417	7.359	−18.378	−19.165	12.289
4	0.32	7.276	17.372	39.506	19.858	41.939
5	0.16	4.82	38.323	37.113	23.744	34.063
6	0.351	11.842	23.118	14.725	11.616	9.516
7	0.243	5.173	17.515	14.435	123.101	79.489
8	−0.19	−10.912	8.236	−2.746	−7.439	−10.502
9	0.173	7.543	23.978	17.122	21.318	25.701
10	0.367	9.352	16.048	55.621	27.861	18.918

数据来源：梅长林，范金城．数据分析方法［M］．北京：高等教育出版社，2006：123.

解： 设原始数据矩阵为

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} \quad (m=10,n=6)$$

（1）利用变异系数法建立权向量

$$w_j = v_j / \sum_{j=1}^{6} v_j$$

其中 $v_j = s_j / |\overline{x}_j|$，$s_j$ 与 $\overline{x}_j$ 分别为第 j 项指标的标准差和均值。

$$w = (0.135\,0, 0.698\,8, 0.014\,9, 0.061\,7, 0.062\,5, 0.027\,0)$$

（2）建立理想方案

$$u = (u_1^0, u_2^0, \cdots, u_6^0)$$

其中 $u_j^0 = \max\limits_{1\leqslant i\leqslant 10}\{x_{ij}\}, j = 1,2,\cdots,6$。

（3）建立相对偏差模糊矩阵 R

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix} \quad (m = 10, n = 6)$$

其中 $r_{ij} = \dfrac{|x_{ij} - u_i^0|}{\max\limits_{1\leqslant i\leqslant 10}\{x_{ij}\} - \min\limits_{0\leqslant i\leqslant 10}\{x_{ij}\}}$，利用 MATLAB 软件得到

$$R = \begin{pmatrix} 0.313\,4 & 0 & 0 & 1.000\,0 & 1.000\,0 & 0.785\,3 \\ 0.461\,1 & 0.380\,9 & 0.820\,3 & 0.701\,8 & 0.784\,9 & 0.934\,7 \\ 1.000\,0 & 1.000\,0 & 1.000\,0 & 0.775\,4 & 0.873\,2 & 0.746\,7 \\ 0.042\,6 & 0.218\,9 & 0.799\,5 & 0.168\,9 & 0.633\,7 & 0.417\,3 \\ 0.187\,5 & 0.246\,4 & 0.380\,1 & 0.193\,9 & 0.609\,9 & 0.504\,8 \\ 0.014\,5 & 0.167\,7 & 0.684\,5 & 0.428\,5 & 0.684\,3 & 0.777\,6 \\ 0.112\,3 & 0.242\,5 & 0.796\,7 & 0.431\,6 & 0 & 0 \\ 0.504\,5 & 0.422\,7 & 0.982\,4 & 0.611\,6 & 0.801\,3 & 1.000\,0 \\ 0.175\,7 & 0.215\,9 & 0.667\,3 & 0.403\,4 & 0.624\,8 & 0.597\,7 \\ 0 & 0.195\,6 & 0.826\,1 & 0 & 0.584\,6 & 0.673\,1 \end{pmatrix}$$

（4）建立综合评价模型

$$D_i = \sum_{j=1}^{6} r_{ij} w_j \quad (i = 1,2,\cdots,10)$$

评价准则为：若 $D_i < D_j$，则第 i 家上市公司的业绩优于第 j 家上市公司的业绩。

经计算可得：

$$D_1 = 0.187\,8, D_2 = 0.458\,3, D_3 = 0.971\,4, D_4 = 0.232\,0, D_5 = 0.266\,9$$

$$D_6 = 0.219\,6, D_7 = 0.223\,1, D_8 = 0.493\,1, D_9 = 0.264\,7, D_{10} = 0.203\,8$$

根据评价准则可得各公司排名如表 2-10 所示。

表 2-10　10 家上市公司的综合排名

编　号	1	2	3	4	5	6	7	8	9	10
排　名	1	8	10	5	7	3	4	9	6	2

说明　如果采取不同的方法建立权向量，或者不同的方法得到相对优属度矩阵，评价的结果会有所不同。

MATLAB 程序如下：

```
clear
X=[0.021    26.806    57.311    -39.815    -39.815    8.819
  -0.142    -7.179    16.335    -11.359     -4.766   -4.626
  -0.737   -62.417     7.359    -18.378    -19.165   12.289
    0.32     7.276    17.372     39.506     19.858   41.939
    0.16      4.82    38.323     37.113     23.744   34.063
   0.351    11.842    23.118     14.725     11.616    9.516
   0.243     5.173    17.515     14.435    123.101   79.489
```

```
  - 0.19   - 10.912    8.236    - 2.746    - 7.439 - 10.502
    0.173    7.543   23.978     17.122     21.318   25.701
    0.367    9.352   16.048     55.621     27.861   18.918];   % 输入原始数据
m= mean(X);                                                   % 计算各指标均值
s= std(X);                                                    % 计算各指标标准差
v= s./abs(m);                                                 % 计算各指标变异系数
w= v/sum(v);                                                  % 计算各指标权重
R= abs(X- ones(10,1) * max(X))./[ones(10,1) * range(X)];      % 相对偏差矩阵
D= R * w';                                                    % 计算综合评价值
[F1,t1]= sort(D);                                             % 综合评价值排序
[F2,t2]= sort(t1)                                             % t2 输出上市公司排名
```

习 题 2

1. 已知样本数据为

$$1,3,4,2,9,6,7,8,11,2.5,3,10$$

（1）求该数据的中位数；（2）该数据的顺序统计量；（3）写出上述计算的 MATLAB 实现程序。

2. 设有数据 $(x_1,x_2,\cdots,x_n)$，用 MATLAB 编写 3 种不同程序，均能实现计算 $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}$。

3. 设矩阵 A 表示某球队参加 5 场比赛的技术统计数据

$$A=(a_1 \quad a_2 \quad \cdots \quad a_6)=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{16} \\ a_{21} & a_{22} & \cdots & a_{26} \\ \vdots & \vdots & & \vdots \\ a_{51} & a_{52} & \cdots & a_{56} \end{pmatrix}$$

其中，a_1 表示投篮命中率，a_2 表示罚球命中率，a_3 表示后场篮板球，a_4 表示失误次数，a_5 表示抢断次数，a_6 表示盖帽次数。（1）指标 a_i 中哪些是效益型指标？哪些是成本型指标？写出统一趋势化的计算公式。（2）写出 MATLAB 中的计算程序。

4. 安徽省 1990—2004 年万元工业 GDP 废气排放量、废水排放量、固体废物排放量以及用于污染治理的投入经费比重见表 2-11，解决以下问题：（1）计算各指标的均值、方差、标准差、变异系数以及相关系数矩阵；（2）计算各指标的偏度、峰度、三均值以及极差；（3）做出各指标数据直方图并检验该数据是否服从正态分布？若不服从正态分布，利用 box-cox 变换以后给出该数据的密度函数；（4）上网查找 1990—2004 年江苏省万元工业 GDP 废气排放量，安徽省与江苏省的废气排放量是否服从同样的分布？

表 2-11 废气、废水、固体废物排放量及污染治理的投入经费占 GDP 比重

年 份	万元工业 GDP 废气排放量（m^3）	万元工业 GDP 固体废物排放量（kg）	万元工业 GDP 废水排放量（kg）	环境污染治理投资占 GDP 比重（%）
1990	104 254.40	519.48	441.65	0.18
1991	94 415.00	476.97	398.19	0.26
1992	89 317.41	119.45	332.14	0.23
1993	63 012.42	67.93	203.91	0.20
1994	45 435.04	7.86	128.20	0.17
1995	46 383.42	12.45	113.39	0.22
1996	39 874.19	13.24	87.12	0.15

（续）

年　份	万元工业 GDP 废气排放量（m^3）	万元工业 GDP 固体废物排放量（kg）	万元工业 GDP 废水排放量（kg）	环境污染治理投资占 GDP 比重（%）
1997	38 412.85	37.97	76.98	0.21
1998	35 270.79	45.36	59.68	0.11
1999	35 200.76	34.93	60.82	0.15
2000	35 848.97	1.82	57.35	0.19
2001	40 348.43	1.17	53.06	0.11
2002	40 392.96	0.16	50.96	0.12
2003	37 237.13	0.05	43.94	0.15
2004	34 176.27	0.06	36.90	0.13

5. 利用 MATLAB 软件生成均值向量为（3，2），协方差矩阵为 $\begin{bmatrix} 1 & 1.5 \\ 1.5 & 4 \end{bmatrix}$ 的二元正态分布的随机数，并给出作散点图以及密度函数曲面图的程序。

实验 1　数据统计量及其分布检验

实验目的

1. 熟练掌握利用 MATLAB 软件计算均值、方差、协方差、相关系数、标准差与变异系数、偏度与峰度、中位数、分位数、三均值、四分位极差与极差。

2. 熟练掌握 jbtest 与 lillietest 关于一元数据的正态性检验。

3. 掌握统计作图方法。

4. 掌握多维数据的数字特征与相关矩阵的处理方法。

实验数据与内容

1949—1990 年我国洪涝灾害统计数据如表 2-12 所示，解决以下问题：（1）计算各项指标的平均值、标准差、变异系数、三均值、偏度与峰度；（2）各项指标是否服从正态分布？若服从正态分布，计算概率为 1%时的受灾面积、受灾人口及直接经济损失；若不服从正态分布，利用 box-cox 变换将数据进行变换，对变换后的数据进行相应的分析。

表 2-12　我国洪涝灾害统计数据

年　份	受灾面积	受灾人口	直接经济损失	年　份	受灾面积	受灾人口	直接经济损失
1949	928.2	2006	190 300	1957	808.27	870	45 708.41
1950	656	1928	12 028.87	1958	428	1132	14 692
1951	417	601	12 614.71	1959	481	845	25 746
1952	279.4	1059	23 339.56	1960	1016	682	58 179.59
1953	741	812	10 897.38	1961	887	1867	26 172.85
1954	1613	3937	209 300	1962	981	1501	53 865.8
1955	525	407	13 061.56	1963	1407	2757	629 755.2
1956	1438	2576	326 801.7	1964	1493	1561	31 458.73

（续）

年　份	受灾面积	受灾人口	直接经济损失	年　份	受灾面积	受灾人口	直接经济损失
1965	559	683	23 751.14	1978	285	2130	26 155.93
1966	251	1079	68 286.03	1979	676	2191	54 798.1
1967	170.89	575	14 286.03	1980	915	4106	90 339.39
1968	224.34	372	8232.32	1981	862	4560	335 319.3
1969	463.18	1252	23 293.55	1982	836	4499	120 239.5
1970	313	305	17 424.71	1983	1216	5294	221 760.3
1971	399	618	15 312.09	1984	1069	nan	1530
1972	408	1608	21 804	1985	1 419.73	1294	470 282
1973	624	1746	14 378.77	1986	915.53	321	703 600
1974	640	1988	35 974.6	1987	868.6	2105	246 253.3
1975	682	1208	1 000 000	1988	1 194.93	3522	803 387.8
1976	420	2589	26 163.63	1989	1 132.8	nan	233 000
1977	910	1872	60 604.77	1990	1 180.4	7611	1 591 968

注：表中 nan 表示数据缺失。

第3章 回归分析

回归分析是最常用的数据分析方法之一。它是根据已得的试验结果以及以往的经验来建立统计模型，并研究变量间的相关关系，建立起变量之间关系的近似表达式（即经验公式），并由此对相应的变量进行预测和控制等。本章将介绍一元回归模型、非线性回归模型、多元线性回归与逐步回归等内容。

3.1 一元回归模型

3.1.1 一元线性回归模型

1. 一元线性回归的基本概念

设 Y 是一个可观测的随机变量，它受到一个非随机变量因素 x 和随机误差 ε 的影响。若 Y 与 x 有如下线性关系：

$$Y=\beta_0+\beta_1 x+\varepsilon \tag{3.1.1}$$

且 ε 的均值 $E(\varepsilon)=0$，方差 $\mathrm{var}(\varepsilon)=\sigma^2(\sigma>0)$，其中 β_0、β_1 是固定的未知参数，称为回归系数，Y 称为因变量，x 称为自变量，则称式（3.1.1）为一元线性回归模型。

对于实际问题，要建立回归方程，首先要确定能否建立线性回归模型，其次确定如何对模型中未知参数 β_0、β_1 进行估计。

通常，我们首先对总体（x，Y）进行 n 次独立观测，获得 n 组数据（称为样本观测值）：

$$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$$

然后在直角坐标系 xoy 中画出数据点 $(x_i,y_i)(i=1,2,\cdots,n)$，该图形称为数据的散点图。如果这些点大致地位于同一条直线的附近，或者说，散点图呈现线性形状，则认为 Y 与 x 之间的关系符合式（3.1.1）。此时，利用最小二乘法可以得到回归模型参数 β_0、β_1 的最小二乘估计 $\hat{\beta}_0$、$\hat{\beta}_1$，估计公式为：

$$\begin{cases}\hat{\beta}_0=\bar{y}-\bar{x}\hat{\beta}_1\\ \hat{\beta}_1=\dfrac{L_{xy}}{L_{xx}}\end{cases} \tag{3.1.2}$$

其中，$\bar{x}=\dfrac{1}{n}\sum_{i=1}^{n}x_i$，$\bar{y}=\dfrac{1}{n}\sum_{i=1}^{n}y_i$，$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$，$L_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$。

于是建立经验公式模型：

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x \tag{3.1.3}$$

一元线性回归分析的主要任务：一是利用样本观测值对回归系数 β_0、β_1 和 σ 做点估计；二是对方程的线性关系（即 β_1）做显著性检验；三是在 $x=x_0$ 处对 Y 做预测等。

下面举例说明如何建立经验公式（3.1.3）。

例 3.1.1 近 10 年来，某市社会商品零售总额与职工工资总额（单位：亿元）的数据见表 3-1。建立社会商品零售总额与职工工资总额数据的回归模型。

表 3-1 商品零售总额与职工工资总额 （单位：亿元）

职工工资总额	23.8	27.6	31.6	32.4	33.7	34.9	43.2	52.8	63.8	73.4
商品零售总额	41.4	51.8	61.7	67.9	68.7	77.5	95.9	137.4	155.0	175.0

解：编写程序如下：

```
% 首先输入数据
x=[23.80,27.60,31.60,32.40,33.70,34.90,43.20,52.80,63.80,73.40];
y=[41.4,51.8,61.70,67.90,68.70,77.50,95.90,137.40,155.0,175.0];
% 然后作散点图
plot(x,y,'*')                    % 作散点图
xlabel('x(职工工资总额)')          % 横坐标名
ylabel('y(商品零售总额)')          % 纵坐标名
```

运行后图形如图 3-1 所示。由于图中的数据点大致地位于同一条直线上，因此可建立一元线性回归模型。程序如下：

```
% 计算最佳参数
Lxx=sum((x-mean(x)).^2);
Lxy=sum((x-mean(x)).*(y-mean(y)));
b1=Lxy/Lxx;
b0=mean(y)-b1*mean(x);
```

运行后得到：

```
b1=2.7991,b0=-23.5493
```

所以，回归模型为：

$$\hat{y} = 2.7991x - 23.5493$$

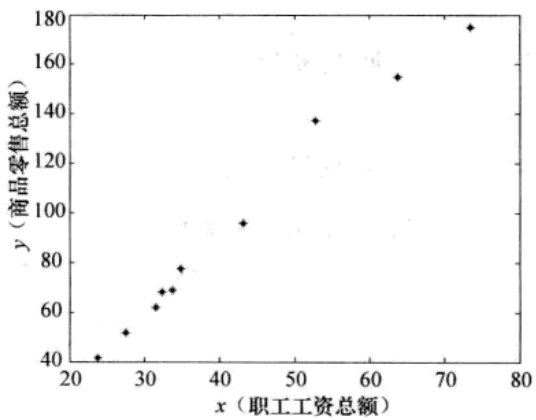


图 3-1 商品零售总额与职工工资总额数据的散点图

该模型表明职工工资总额每增加 1 亿元，社会商品零售总额将增加 2.80 亿元。

2. 一元多项式回归模型

在一元回归模型中，如果变量 y 与 x 的关系是 n 次多项式，即

$$y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 + \varepsilon \tag{3.1.4}$$

其中，ε 是随机误差，服从正态分布 $N(0, \sigma^2)$，$a_0, a_1, \cdots, a_n$ 为回归系数，则称式（3.1.4）为一元多项式回归模型。

（1）多项式曲线拟合

在 MATLAB 7.0 的统计工具箱中，有多项式曲线拟合的命令 polyfit，其调用格式有以下三种：

```
p=polyfit(x,y,n)
[p,S]=polyfit(x,y,n)
[p,S,mu]=polyfit(x,y,n)
```

其中，输入 x，y 分别为自变量与因变量的样本观测数据向量；n 是多项式的阶数，对于一元线性回归取 $n=1$；输出 p 是按照降幂排列的多项式的系数向量；S 是一个矩阵，用于估计预测误差或供 MATLAB 的其他函数（如 polyconf、polyval 等）的调用；mu 是一个向量给出自变量的均值与标准差。

下面利用 polyfit 命令，给出例 3.1.1 的另一种解法。

解：编写程序如下：

```
% 首先输入数据
x= [23.80,27.60,31.60,32.40,33.70,34.90,43.20,52.80,63.80,73.40];
y= [41.4,51.8,61.70,67.90,68.70,77.50,95.90,137.40,155.0,175.0];
% 然后调用一元回归命令
p= polyfit(x,y,1)              % 注意取 n= 1
```

运行得到：

```
p=
   2.7991   - 23.5493
```

即回归模型为：

$$\hat{y} = 2.799\,1x - 23.549\,3$$

比较这两种解法，第二种解法的程序要简洁些。

例 3.1.2 某种合金的主要成分为 A、B 两种金属，经过试验发现：这两种金属成分之和 x 与合金的膨胀系数 y 的关系见表 3-2，建立描述这种关系的数学表达式。

表 3-2 合金的膨胀系数表

x	37	37.5	38	38.5	39	39.5	40	40.5	41	41.5	42	42.5	43
y	3.4	3	3	2.27	2.1	1.83	1.53	1.7	1.8	1.9	2.35	2.54	2.9

解：编写程序如下：

```
% 首先作出散点图
x= 37:0.5:43;
y= [3.4,3,3,2.27,2.1,1.83,1.53,1.7,1.8,1.9,2.35,2.54,2.9];
plot(x,y,'*')
xlabel('x(两种合金之和)')          % 横坐标名
ylabel('y(合金膨胀系数)')          % 纵坐标名
```

图形如图 3-2 所示。由于散点图呈现抛物线形状，因此选择二次函数曲线进行拟合，命令如下：

```
p= polyfit(x,y,2)          % 注意取 n= 2
```

运行得到回归系数：

```
p=
   0.1660   - 13.3866   271.6231
```

即二次回归模型为：

$$\hat{y} = 0.166x^2 - 13.386\,6x + 271.623\,1$$

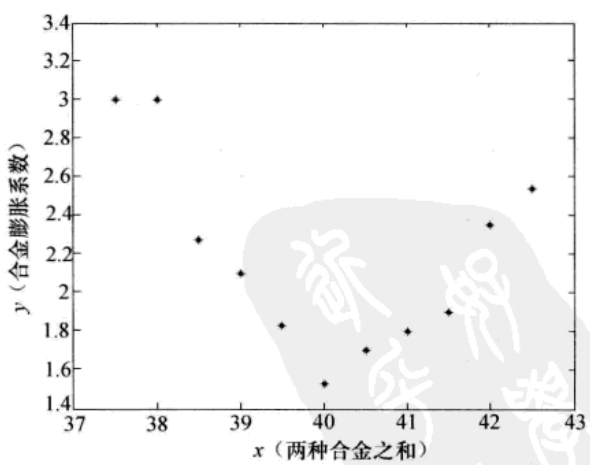


图 3-2 合金系数散点图

（2）多项式回归模型的预测及其置信区间

在 MATLAB 7.0 的统计工具箱中，有多项式曲线拟合预测的命令 polyval，其调用格式有以下两种：

```
Y= polyval(p,x0)
[Y,Delta]= polyconf(p,x0,S,alpha)
```

其中，输入 p、S 是多项式拟合命令 [p，S]=polyfit（x，y，n）的输出，$x0$ 是要预测的自变量的值；输出 Y 是 polyfit 所得的回归多项式在 x 处的预测值，如果输入数据的误差相互独立，

且方差为常数，则 Y±Delta 至少包含 50%的预测值；alpha 默认值为 0.05。

（3）多项式回归的 GUI 界面命令

在 MATLAB 7.0 的统计工具箱中，有一个调用多项式回归的 GUI 界面命令 polytool，其典型调用格式为：

```
polytool(x,y,n,alpha)
```

其中，输入 x、y 分别为自变量与因变量的样本观测数据向量；n 是多项式的阶数；置信度为(1－alpha)%，alpha 默认值为 0.05。

该命令可以绘出总体拟合图形以及（1－alpha）上、下置信区间的直线（屏幕上显示为红色）。此外，用鼠标拖动图中纵向虚线，就可以显示出对于不同的自变量数值所对应的预测状况。与此同时，图形左端数值框中会随着自变量的变化而得到预测数值以及（1－alpha）置信区间长度一半的数值。

下面举例说明上述多项式回归的 MATLAB 命令的应用方法。

例 3.1.3 为了分析 X 射线的杀菌作用，用 200 千伏的 X 射线来照射细菌，每次照射 6 分钟，用平板计数法估计尚存活的细菌数。照射次数记为 t，照射后的细菌数为 y 见表 3-3。试求：

①给出 y 与 t 的二次回归模型。

②在同一坐标系内作出原始数据与拟合结果的散点图。

③预测 $t=16$ 时残留的细菌数。

④根据问题的实际意义，你认为选择多项式函数是否合适？

表 3-3 X 射线照射次数与残留细菌数

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	352	211	197	160	142	106	104	60	56	38	36	32	21	19	15

数据来源：http//www.ilr.cornell.edu/～hadi/RABE.

解：编写程序如下：

```
% 输入原始数据
t= 1:15;
y= [352,211,197,160,142,106,104,60,56,38,36,32,21,19,15];
p= polyfit(t,y,2);          % 作二次多项式回归
y1= polyval(p,t);           % 模型估计与作图
plot(t,y,'- * ',t,y1,'- o');
legend('原始数据','二次函数')
xlabel('t(照射次数)')
ylabel('y(残留细菌数)')
t0= 16;
yc1= polyconf(p,t0)       % 预测 t0= 16 时残留的细菌数
```

运行结果为：

```
p=
    1.9897   - 51.1394   347.8967
yc1=
     39.0396
```

即二次回归模型为：

$$y_1 = 1.9897t^2 - 51.1394t + 347.8967$$

原始数据与拟合结果的散点图如图 3-3 所示，从图形可知拟合效果较好。即照射 16 次后，用二次函数计算出细菌残留数为 39.0396，显然与实际不符。

在命令窗口中输入调用多项式回归的 GUI 界面命令：

```
polytool(t,y,2)
```

则打开如图 3-4 所示的 GUI 窗口，显示 X 射线照射次数与残留细菌数拟合交互图。

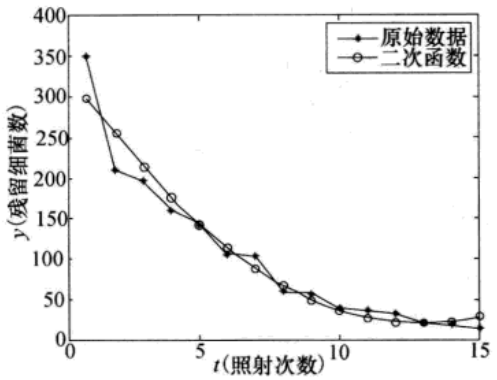


图 3-3 原始数据与拟合结果的散点图

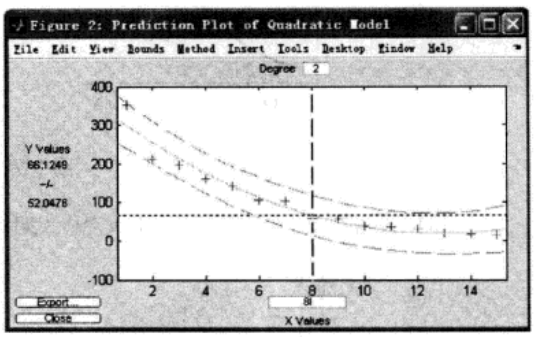


图 3-4 X 射线照射次数与残留细菌数拟合交互图

根据实际问题的意义可知：尽管二次多项式拟合效果较好，但是用于预测并不理想。因此如何根据原始数据散点图的规律，选择适当的回归曲线是非常重要的，这样就有必要研究非线性回归模型。

3.1.2 一元非线性回归模型

1. 非线性曲线选择

为了便于正确地选择合适的函数进行回归分析建模，我们给出如下通常选择的 6 类曲线：

1）双曲线 $\frac{1}{y}=a+\frac{b}{x}$（如图 3-5 所示）。

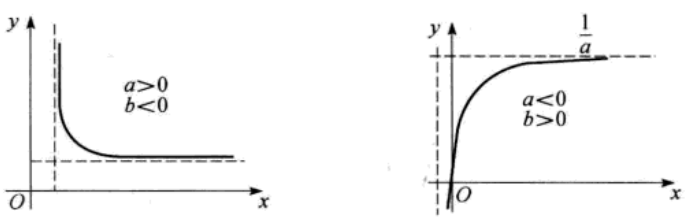


图 3-5 双曲线

2）幂函数曲线 $y=ax^b$，其中 $x>0$，$a>0$（如图 3-6 所示）。

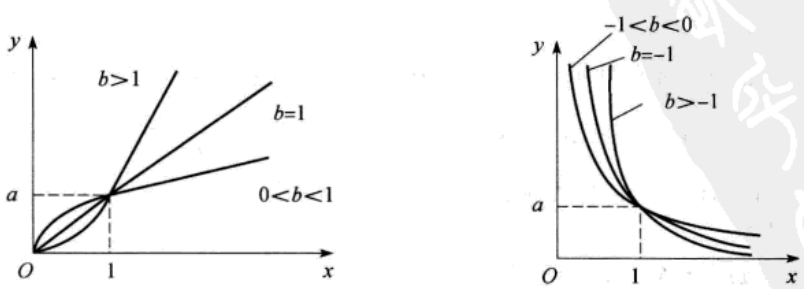


图 3-6 幂函数曲线

3）指数曲线 $y=ae^{bx}$，其中参数 $a>0$（如图 3-7 所示）。

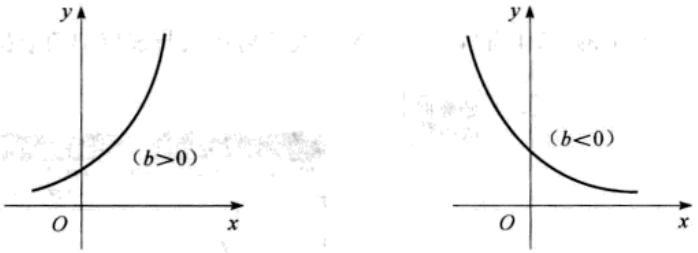


图 3-7 指数曲线

4）倒指数曲线 $y=ae^{b/x}$，其中 $a>0$（如图 3-8 所示）。

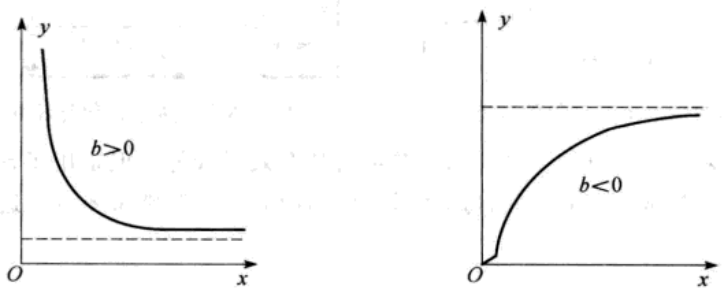


图 3-8 倒指数曲线

5）对数曲线 $y=a+b\ln x$（如图 3-9 所示）。

6）S 型曲线 $y=\dfrac{1}{a+be^{-x}}$，其中 $ab>0$（如图 3-10 所示）。

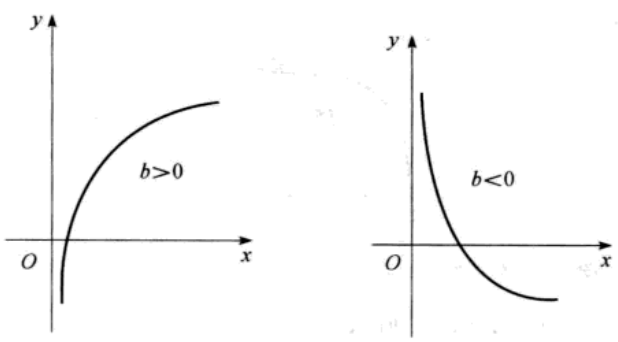


图 3-9 对数曲线

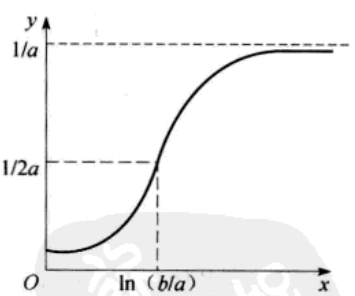


图 3-10 S 型曲线

非线性回归建模通常有两种方法：一是通过适当的变换转化为线性回归模型，例如双曲线模型 $\dfrac{1}{y}=a+\dfrac{b}{x}$（如图 3-5 所示），如果作变换 $y'=\dfrac{1}{y}$，$x'=\dfrac{1}{x}$，则有 $y'=a+bx'$，此时 x'、y'就是一阶线性回归模型；如果无法实现线性化，可以利用最小二乘法直接建立非线性回归模型，求解最佳参数。

2. 非线性回归的 MATLAB 命令

MATLAB 统计工具箱中实现非线性回归的命令有 nlinfit、nlpredci、nlparci 和 nlintool。下面逐一介绍调用格式。

1）非线性拟合命令为 nlinfit，其调用格式为：

```
[beta,r,J]= nlinfit(x,y,'model',beta0)
```

其中，输入数据 x、y 分别为 $n\times m$ 矩阵和 n 维列向量，对一元非线性回归，x 为 n 维列向量，model 是事先用 M 文件定义的非线性函数，beta0 是回归系数的初值（需要通过解方程组得到），beta 是估计出的最佳回归系数，r 是残差，J 是雅可比（Jacobian）矩阵，它们是估计预测误差需要的数据。通常，可以利用 inline 定义函数 model，方法如下：

```
fun= inline('f(x)','参变量','x')
```

2）非线性回归预测命令为 nlpredci，其调用格式为：

```
ypred= nlpredci(FUN,inputs,beta,r,J)
```

其中，输入参数 beta、r、J 是非线性回归命令 nlinfit 的输出结果，FUN 是拟合函数，inputs 是需要预测的自变量；输出量 ypred 是 inputs 的预测值。

3）非线性回归置信区间命令为 nlparci，其调用格式为：

```
ci= nlparci(beta,r,J,alpha)
```

其中，输入参数 beta、r、J 是非线性拟合命令 nlinfit 输出的结果；输出 ci 是一个矩阵，每一行分别为每个参数的（1－alpha）%的置信区间，alpha 默认值为 0.05。

4）非线性回归交互命令 nlintool，其典型调用格式为：

```
nlintool(x,y,fun,beta0)
```

其中参数 x、y、fun、beta0 与命令 nlinfit 中的参数含义相同。GUI 界面与多项式回归的命令 polytool 的界面相似，此处不再重述。

下面举例说明上述非线性回归的 MATLAB 命令的应用方法。

例 3.1.4 在 M 文件中建立函数 $y=a(1-be^{-\alpha})$，其中 a、b、c 为待定的参数。

解：程序如下：

```
fun= inline('b(1) * (1- b(2) * exp(- b(3) * x))','b','x');
```

此处，将 b 看成参变量，$b(1)$、$b(2)$、$b(3)$ 为其分量。

例 3.1.5 炼钢厂出钢时所用盛钢水的钢包，由于钢水对耐火材料的侵蚀，容积不断增大，我们希望找出使用次数与增大容积之间的函数关系。实验数据见表 3-4。

（1）建立非线性回归模型 $\frac{1}{y}=a+\frac{b}{x}$。

（2）预测钢包使用 $x0=17$ 次后增大的容积 $y0$。

（3）计算回归模型参数的置信度为 95%的置信区间。

表 3-4 钢包使用次数与增大容积

使用次数（x）	2	3	4	5	6	7	8	9
增大容积（y）	6.42	8.2	9.58	9.5	9.7	10	9.93	9.99
使用次数（x）	10	11	12	13	14	15	16	
增大容积（y）	10.49	10.59	10.6	10.8	10.6	10.9	10.76	

解：MATLAB 程序如下：

```
x= [2:16];
y= [6.42,8.2,9.58,9.5,9.7,10,9.93,9.99,10.49,10.59,10.6,10.8,10.6,10.9,10.76];
% 建立非线性双曲线回归模型
b0= [0.084,0.1436];                                    % 初始参数值
fun= inline('x./(b(1) * x+ b(2))','b','x');
[beta,r,J]= nlinfit(x,y,fun,b0);
```

```
beta                                    % 输出最佳参数
yl= x./(0.0845 * x+ 0.1152);            % 拟合曲线
plot(x,y,' * ',x,yl,'- or')
legend('原始数据','拟合曲线')
```

初始值要先计算后，才能得到上面程序中的 b0，由于确定两个参数值，因此我们选择已知数据中的两点（2，6.42）和（16，10.76）代入设定方程，得到方程组

$$\begin{cases} 6.42=\dfrac{2}{2a+b} \\ 10.76=\dfrac{16}{16a+b} \end{cases} \Rightarrow \begin{cases} 6.42\ (2a+b)\ =2 \\ 10.76\ (16a+b)\ =16 \end{cases}$$

上述方程组有两种解法：手工方法与 MATLAB 方法。下面用 MATLAB 方法解方程组：

```
[a,b]= solve('6.42 * (2 * a+ b)= 2','
10.76 * (16 * a+ b)= 16')
```

解得

$a=$

.839 615 977 023 474 504 626 573 556 150 04e−1≈0.084

$b=$

.143 603 284 346 083 915 274 062 235 810 49≈0.143 6

钢包使用次数与增大容积的非线性拟合图如图 3-11 所示。

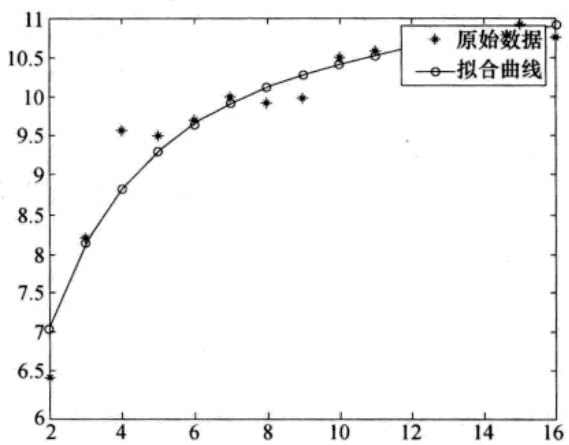


图 3-11 钢包使用次数与增大容积的非线性拟合图

在例 3.1.5 中，为预测钢包使用 17 次后增大的容积，可在上面执行的程序中继续输入命令

```
ypred= nlpredci(fun,17,beta,r,J)
```

得到：

```
ypred= 10.9599
```

即钢包使用 17 次后增大的容积为 10.959 9。

求回归模型参数的置信度为 95%的置信区间，只要继续添加程序

```
ci= nlparci(beta,r,J)
```

运行后得到：

```
ci=
    0.0814    0.0876
    0.0934    0.1370
```

即回归模型 $y=\dfrac{x}{ax+b}$ 中参数 a、b 的置信度为 95%的置信区间分别为 [0.081 4，0.087 6] 与 [0.093 4，0.137 0]。我们求出的最佳参数分别为 a=0.084 5 和 b=0.115 2，均属于上述置信区间。

图 3-12 给出了例 3.1.5 钢包使用次数与增大容积的非线性拟合的交互图形，图中的圆圈是实验的原始数据点，两条虚线为 95%上、下置信区间的曲线（屏幕上显示为红色），中间的实线（屏幕上显示为绿色）是回归模型曲线，纵向的蓝色虚线显示自变量为 8.950 2 时，横向虚线对应的预测值为 10.273 4。

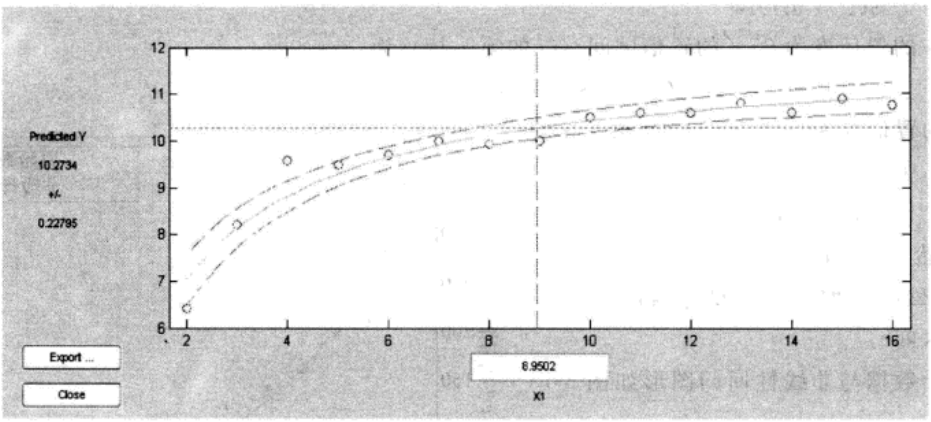


图 3-12 钢包使用次数与增大容积的非线性拟合交互图

例 3.1.6 对例 3.1.3 进行非线性回归，并预测照射 16 次后细菌残留数目，给出模型参数的置信度为 95%的置信区间，并绘出模型交互图形。

解：我们选取函数 $y=ae^{bt}$ 进行非线性回归，该方程的两个参数具有简单的物理解释，a 表示实验开始时的细菌数目，b 表示细菌死亡（或衰变）的速率。

MATLAB 程序如下：

```
t= 1:15;
y= [352  211  197  160  142  106  104  60  56  38  36  32  21  19  15];
fun= inline('b(1) * exp(b(2) * t)','b','t')      % 非线性函数
beta0= [148,- 0.2];                              % 参数初始值
[beta,r,J]= nlinfit(t,y,fun,beta0);              % 非线性拟合
beta                                             % 输出最佳参数
y1= nlpredci(fun,t,beta,r,J);                    % 模型数值计算
plot(t,y,'* ',t,y1,'- or'),
legend('原始数据','非线性回归')
xlabel('t(照射次数)')
ylabel('y(残留细菌数)')
ypred= nlpredci(fun,16,beta,r,J)                 % 预测残留细菌数
ci= nlparci(beta,r,J)                            % 参数 95% 区间估计
nlintool(t,y,fun,beta0)                          % 作出交互图形
```

运行结果如下：

```
beta=
    400.0904   - 0.2240
```

即最佳参数为：$a=400.0904$，$b=-0.2240$。故非线性回归模型为

$$y = 400.0904e^{-0.224t}$$

```
ypred=
    11.1014
```

即照射 16 次后细菌残留数目为 11.1014，该预测符合实际，显然比例 3.1.3 中多项式回归的结果合理。

```
ci=
   355.2481  444.9326
```

```
    - 0.2561   - 0.1919
```

即参数 a 的置信度为 95%的置信区间（ci 的第一行）为：

$$[355.248\,1, 444.932\,6]$$

参数 b 的置信度为 95%的置信区间（ci 的第二行）为：

$$[-0.256\,1, -0.191\,9]$$

显然，最佳参数 $a=400.090\,4$，$b=-0.224\,0$，均属于各自置信度为 95%的置信区间。

原始数据与非线性回归图形如图 3-13 所示。

从图 3-14 可以看出：圆圈为原始数据；两条虚线（屏幕上显示红色）是置信区间曲线；两条虚线内的实线（屏幕上显示绿色）是回归模型曲线；纵向虚线指示照射 8 次，此时对应的水平虚线表示模型得到的残留细菌数为 66.645 1。

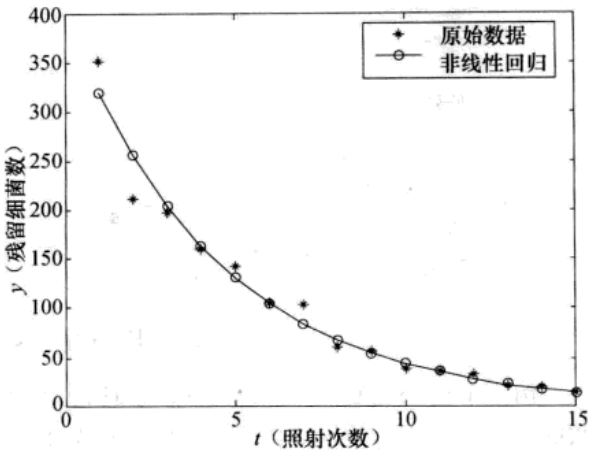


图 3-13　原始数据与非线性回归图形

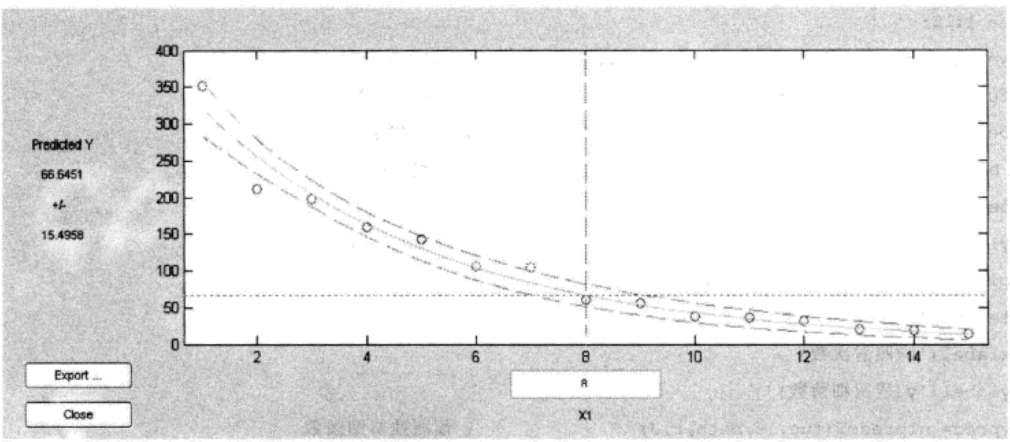


图 3-14　原始数据与非线性回归 GUI 图形

3.1.3　一元回归建模实例

例 3.1.7　在四川白鹅的生产性能研究中，得到如下一组关于雏鹅重与 70 日龄重的数据，试建立 70 日龄重（y）与雏鹅重（x）的直线回归方程，并计算模型误差平方和以及可决系数。当雏鹅重分别为 85、95、115 时，预测其 70 日龄重和置信区间。

表 3-5　雏鹅重与 70 日龄重测定结果　（单位：g）

编　号	1	2	3	4	5	6	7	8	9	10	11	12
雏鹅重（x）	80	86	98	90	120	102	95	83	113	105	110	100
70 日龄重（y）	2 350	2 400	2 720	2 500	3 150	2 680	2 630	2 400	3 080	2 920	2 960	2 860

解：(1) 作散点图。以雏鹅重（x）为横坐标，70日龄重（y）为纵坐标作散点图，如图3-15所示。

在MATLAB命令窗口中输入：

```
x= [80 86 98 90 120 102 95 83 113 105 110
100]';                          % 雏鹅重
y= [2350 2400 2720 2500 3150  2680 2630
2400 3080 2920 2960 2860]';  % 70日龄重
plot(x,y,'* ')                  % 作散点图
xlabel('x(雏鹅重)')             % 横坐标名
ylabel('y(70日龄重)')           % 纵坐标名
```

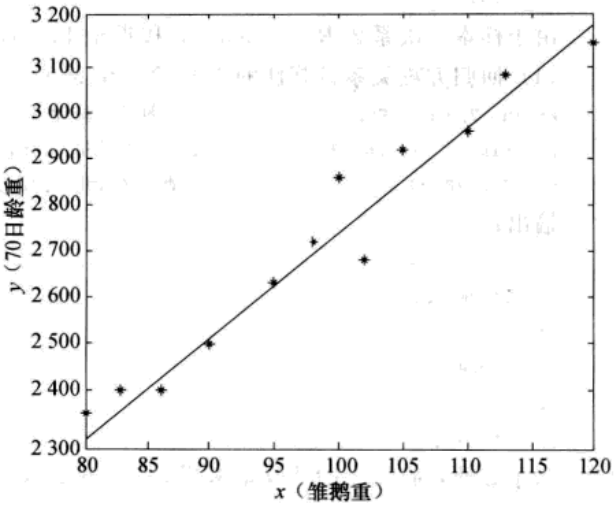


图3-15 四川白鹅的雏鹅重与70日龄重散点图和回归直线图

由图形3-15可见，白鹅的70日龄重与雏鹅重间存在线性关系，且70日龄重随雏鹅重的增大而增大。因此，y 与 x 符合一元线性回归模型。

(2) 建立直线回归方程。在MATLAB中调用命令polyfit，从而求出参数 β_0、β_1 的最小二乘估计。在MATLAB命令窗口中继续输入：

```
n= size(x,1)                    % 计算样本容量
[p,s]= polyfit(x,y,1);          % 调用命令 polyfit 计算回归参数
y1= polyval(p,x);               % 计算回归模型的函数值
hold on
plot(x,y1)                      % 作回归方程的图形,结果如图 3-15 所示
p                               % 显示参数的最小二乘估计结果
```

输出：

```
p=
   21.7122  582.1850
```

即参数（β_0，β_1）的最小二乘估计为：

$$\hat{\beta}_0 = 582.185\,0, \hat{\beta}_1 = 21.712\,2$$

所以70日龄重（y）与雏鹅重（x）的直线回归方程为：

$$\hat{y} = 582.185\,0 + 21.712\,2x$$

(3) 误差估计与可决系数。在MATLAB命令窗口中继续输入：

```
TSS= sum((y- mean(y)).^2)       % 计算总离差平方和
RSS= sum((y1- mean(y)).^2)      % 计算回归平方和
ESS= sum((y- y1).^2)            % 计算残差平方和
R2= RSS/TSS;                    % 计算样本可决系数 R2
```

输出：

```
TSS=
     8.314917e+ 005
RSS=
     7.943396e+ 005
ESS=
     3.715217e+ 004
```

```
R2=
    0.9553
```

由于样本可决系数 $R^2=0.9553$，接近于 1，因此模型拟合的效果较好。

（4）回归方程关系显著性的 F 检验。在 MATLAB 命令窗口中继续输入：

```
F= (n- 2) * RSS/ESS              % 计算的 F 统计量
F1= finv(0.95,1,n- 2)            % 查 F 统计量 0.05 的分位数
F2= finv(0.99,1,n- 2)            % 查 F 统计量 0.01 的分位数
```

输出：

```
F=
   2.138e+ 002
F1=
    4.9646
F2=
    10.0442
```

为了方便，将以上的计算结果列成表 3-6 的形式。

表 3-6 四川白鹅 70 日龄重与雏鹅重回归关系方差分析表

类　型	自由度（df）	平方和（SS）	均方和（MS）	F	$F_{0.05}$	$F_{0.01}$
回归	1	794 339.60	794 339.60	213.81**	4.96	10.04
残差	10	37 152.07	3 715.21			
总离差	11	831 491.67				

因为 $F=213.81>F_2=F_{0.01(1,10)}=10.04$，表明四川白鹅 70 日龄重与雏鹅重间存在显著的线性关系。

（5）回归关系显著性的 t 检验。在 MATLAB 命令窗口中继续输入：

```
T= p(2)/sqrt(ESS/(n- 2)) * sqrt(sum((x- mean(x)).^2))% 计算 T 统计量
T1= tinv(0.975,n- 2)                    % t 统计量 0.05 的分位数
T2= tinv(0.995,n- 2)                    % t 统计量 0.01 的分位数
```

输出：

```
T=
   14.622
T1=
    2.228
T2=
    3.169
```

因为 $T=14.62>T2=t_{0.01(10)}=3.169$，否定 H_0，接受 H_1，即四川白鹅 70 日龄重（y）与雏鹅重（x）的回归系数 $\beta_1=21.712\,2$ 是显著的，表明四川白鹅 70 日龄重与雏鹅重间存在显著的线性关系，可用所建立的回归方程进行预测和控制。

（6）预测，程序如下：

```
x1= [85,95,115]';                % 输入自变量
yc= polyval(p,x1)                % 计算预测值
[Y,Delta]= polyconf(p,x1,s);
I1= [Y- Delta,Y+ Delta]          % 置信区间
```

输出：

```
yc=
    2427.72
```

```
    2644.84
    3079.08
I1=
    2279.47      2575.96
    2503.01      2786.67
    2927.55      3230.62
```

所以当雏鹅重分别为85，95，115时，白鹅70日龄重分别为2 427.72，2 644.84，3 079.08；且95%的置信区间分别为：[2 279.47，2 575.96]，[2 503.01，2 786.67]，[2 927.55，3 230.62]。

在程序中加入：

```
polytool(x,y)                          % 交互功能
bar(x,y- y1),                          % 残差图
legend('残差')
h= lillietest(y- y1)                   % 残差正态性检验
```

输出：

```
h=
   0
```

得到交互图形如图3-16所示，可以看出当雏鹅重为100时，模型给出70日龄鹅重为2 753.401 6。

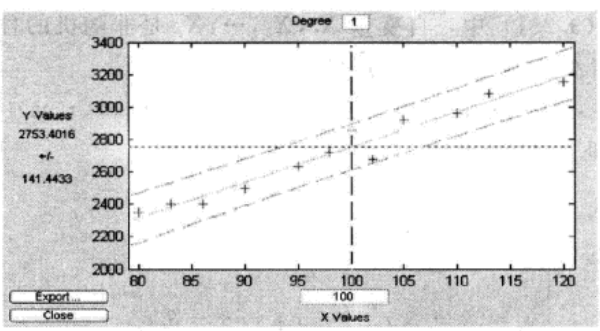


图3-16 四川白鹅70日龄重与雏鹅重线性模型交互图

图3-17为模型残差图，可以看出模型残差没有相关性，正态性检验表明无法拒绝正态分布。

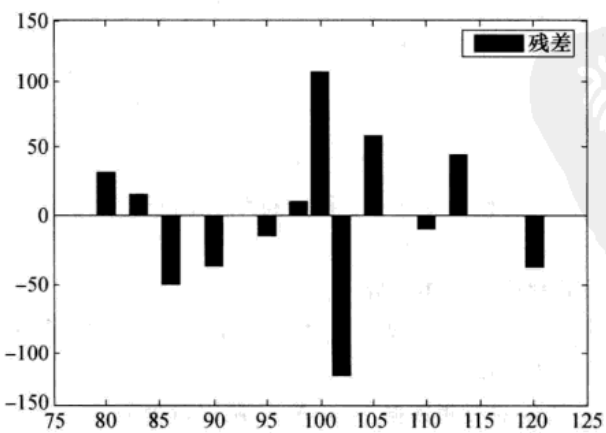


图3-17 模型残差图

说明 求解例 3.1.7 的全部 MATLAB 命令，可以写成一个 M 文件，替换文件中的数据 x，y 就可以适合其他问题的求解。

3.2 多元线性回归模型

上一节介绍的一元回归模型，只能分析两个变量间的相关关系。在很多实际问题中，和某个变量 Y 有关系的变量不止一个，研究一个变量和多个变量之间的定量关系的问题就称为多元回归问题。和建立一元线性回归模型的方法类似，以下我们着重研究多元线性回归模型的建立方法。

3.2.1 多元线性回归模型及其表示

1. 多元线性回归模型的基本概念

设 Y 是一个可观测的随机变量，它受到 $p(p>0)$ 个非随机变量因素 $X_1,X_2,\cdots,X_p$ 和随机误差 ε 的影响。若 Y 与 $X_1,X_2,\cdots,X_p$ 有如下线性关系：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+\varepsilon \tag{3.2.1}$$

其中 $\beta_0,\beta_1,\beta_2,\cdots,\beta_p$ 是固定的未知参数，称为回归系数；ε 是均值为 0、方差为 $\sigma^2(\sigma>0)$ 的随机变量；Y 称为被解释变量；$X_1,X_2,\cdots,X_p$ 称为解释变量。模型 (3.2.1) 称为多元线性回归模型。

由定义，在模型 (3.2.1) 中，自变量 $X_1,X_2,\cdots,X_p$ 是非随机的且可精确观测，随机误差 ε 代表其他随机因素对因变量 Y 产生的影响。

对于总体 $(X_1,X_2,\cdots,X_p;\ Y)$ 的 n 组观测值 $(x_{i1},x_{i2},\cdots,x_{ip};\ y_i)(i=1,2,\cdots,n;\ n>p)$，应满足式 (3.2.1)，即

$$\begin{cases} y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_px_{1p}+\varepsilon_1 \\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_px_{2p}+\varepsilon_2 \\ \qquad\cdots \\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_px_{np}+\varepsilon_n \end{cases} \tag{3.2.2}$$

其中 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立，且设 $\varepsilon_i\sim N(0,\ \sigma^2)(i=1,2,\cdots,n)$，记

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\quad X=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},\quad \beta=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},\quad \varepsilon=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

则模型 (3.2.2) 可用矩阵形式表示为

$$Y=X\beta+\varepsilon \tag{3.2.3}$$

其中 Y 称为观测向量；X 称为设计矩阵；β 称为待估计向量；ε 是不可观测的 n 维随机向量，它的分量相互独立，假定 $\varepsilon\sim N(0,\ \sigma^2I_n)$。

2. 建立多元线性回归建模的基本步骤

1) 对问题进行直观分析，选择因变量与解释变量，作出因变量与各解释变量的散点图，初步设定多元线性回归模型的参数个数。

2) 输入因变量与自变量的观测数据 $(y,\ X)$，调用命令为：

```
[b,bint,r,rint,s]= regress(y,X,alpha)
```

计算参数的估计。

3）调用命令 rcoplot（r，rint），分析数据的异常点情况。

4）作显著性检验，若通过，则对模型作预测。

5）对模型进一步研究，如残差的正态性检验、残差的异方差检验、残差的自相关性检验等。

3.2.2 MATLAB 的回归分析命令

在 MATLAB 7.0 的统计工具箱中，与多元回归模型有关的命令有多个，下面逐一介绍。

1. 多元回归建模命令

多元回归建模命令为 regeress，其调用格式有以下三种：

1）`b= regress(Y,X)`

2）`[b,bint,r,rint,stats]= regress(Y,X)`

3）`[b,bint,r,rint,stats]= regress(Y,X,alpha)`

三种方式的主要区别是输出项参数的多少，第三种方式可称为全参数方式。以第三种为例来说明 regeress 命令的输入与输出参数的含义。

输入参数：输入量 Y 表示模型（3.1.1）中因变量的观测向量 $(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$；$X$ 是一个 $n\times(p+1)$的矩阵，其中第一列元全部是数“1”，第 j 列是自变量 X_j 的观测向量 $(x_{1j},_{2j},\cdots,x_{nj})^{\mathrm{T}}$ $(j=1,2,\cdots,p)$，即

$$Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\quad X=\begin{pmatrix}1&x_{11}&x_{12}&\cdots&x_{1p}\\1&x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&\vdots&\cdots&\vdots\\1&x_{n1}&x_{n2}&\cdots&x_{np}\end{pmatrix}$$

对一元线性回归，取 $p=1$ 即可；alpha 为显著性水平（默认值为 0.05）。

输出参数：输出向量 b 为回归系数估计值，bint 为回归系数的（1－alpha）置信区间；输出向量 r 表示残差列向量，即

$$(y_1-\hat{y}_1,y_2-\hat{y}_2,\cdots,y_n-\hat{y}_n)^{\mathrm{T}}$$

输出量 rint 为模型的残差的（1－alpha）的置信区间；输出量 stats 是用于检验回归模型的统计量，有 4 个分量值：第一个是 R^2，其中 R 是相关系数；第二个是 F 统计量值；第三个是与统计量 F 对应的概率 P，当 $P<$alpha 时拒绝 H_0，即认为线性回归模型有意义；第四个是方差 σ^2 的无偏估计。

例 3.2.1 某销售公司将其连续 18 个月的库存资金额、广告投入、员工薪酬以及销售额四方面的数据作了汇总（见表 3-7）。该公司的管理人员试图根据这些数据找到销售额与其他三个变量之间的关系，以便进行销售额预测并为未来的工作决策提供参考依据。（1）试建立销售额的回归模型；（2）如果未来某月库存资金额为 150 万元，广告投入为 45 万元，员工薪酬为 27 万元，试根据建立的回归模型预测该月的销售额。

表 3-7 某销售公司连续 18 个月的库存资金额、广告投入、员工薪酬、销售额汇总表（单位：万元）

月　份	库存资金额（$x1$）	广告投入（$x2$）	员工薪酬（$x3$）	销售额（y）
1	75.2	30.6	21.1	1 090.4
2	77.6	31.3	21.4	1 133.0
3	80.7	33.9	22.9	1 242.1
4	76.0	29.6	21.4	1 003.2
5	79.5	32.5	21.5	1 283.2
6	81.8	27.9	21.7	1 012.2

（续）

月　份	库存资金额（$x1$）	广告投入（$x2$）	员工薪酬（$x3$）	销售额（y）
7	98.3	24.8	21.5	1 098.8
8	67.7	23.6	21.0	826.3
9	74.0	33.9	22.4	1 003.3
10	151.0	27.7	24.7	1 554.6
11	90.8	45.5	23.2	1 199.0
12	102.3	42.6	24.3	1 483.1
13	115.6	40.0	23.1	1 407.1
14	125.0	45.8	29.1	1 551.3
15	137.8	51.7	24.6	1 601.2
16	175.6	67.2	27.5	2 311.7
17	155.2	65.0	26.5	2 126.7
18	174.3	65.4	26.8	2 256.5

解：为了确定销售额与库存资金额、广告投入、员工薪酬之间的关系，分别作出 y 与 x_1、x_2、x_3 的散点图，若散点图显示它们之间近似线性关系，则可设定 y 与 x_1、x_2、x_3 的关系为三元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

编写程序如下：

```
% 输入数据并作散点图(图 3-18)
A= [75.2    30.6    21.1    1090.4
    77.6    31.3    21.4    1133
    80.7    33.9    22.9    1242.1
    76      29.6    21.4    1003.2
    79.5    32.5    21.5    1283.2
    81.8    27.9    21.7    1012.2
    98.3    24.8    21.5    1098.8
    67.7    23.6    21       826.3
    74      33.9    22.4    1003.3
   151      27.7    24.7    1554.6
    90.8    45.5    23.2    1199
   102.3    42.6    24.3    1483.1
   115.6    40      23.1    1407.1
   125      45.8    29.1    1551.3
   137.8    51.7    24.6    1601.2
   175.6    67.2    27.5    2311.7
   155.2    65      26.5    2126.7
   174.3    65.4    26.8    2256.5];
[m,n]= size(A);
subplot(3,1,1),plot(A(:,1),A(:,4),'+ '),
xlabel('x1(库存资金额)')
ylabel('y(销售额)')
subplot(3,1,2),plot(A(:,2),A(:,4),'* '),
xlabel('x2(广告投入)')
```

```
ylabel('y(销售额)')
subplot(3,1,3),plot(A(:,3),A(:,4),'x'),
xlabel('x3(员工薪酬)')
ylabel('y(销售额)')
```

所得图形如图 3-18 所示，可见销售额 y 与库存资金额、广告投入、员工薪酬具有线性关系，因此可以建立三元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

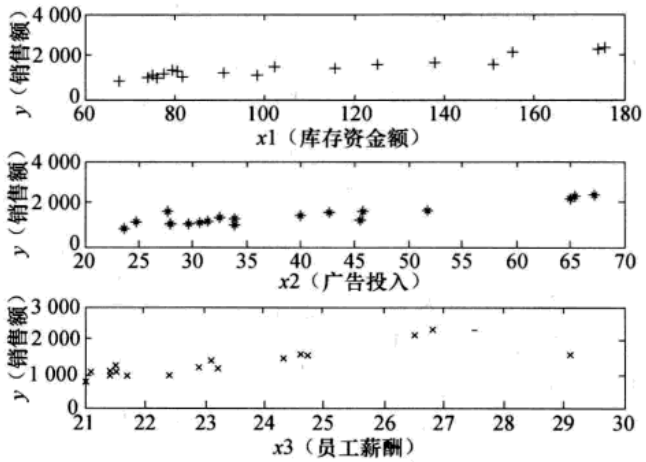


图 3-18　销售额与库存资金额、广告投入、员工薪酬散点图

```
% 调用命令 regress 建立三元线性回归模型
x= [ones(m,1),A(:,1),A(:,2),A(:,3)];
y= A(:,4)
[b,bint,r,rint,stats]= regress(y,x);
b,bint,stats,                                    % 输出结果
```

程序运行结果

```
b=
   162.0632
     7.2739
    13.9575
   - 4.3996
bint=
      - 580.3603   904.4867
         4.3734    10.1743
         7.1649    20.7501
       - 46.7796   37.9805
stats=
        0.9574804050   105.0866520891   0.0000000008   10077.9867891125
```

输出结果说明，b 就是模型中的参数 β_0、β_1、β_2、β_3、因此回归模型为

$$\hat{y} = 162.063\,2 + 7.273\,9x_1 + 13.957\,9x_2 - 4.399\,6x_3$$

bint 的各行分别为参数 β_0、β_1、β_2、β_3 的 95%的置信区间。stats 的第一列表示模型可决系数，第二列为 F 统计量的观测值，第三列得到概率 p=0.000 000 000 8，最后一列为模型的残差平

方和。

由于可决系数 $R^2=0.9575$，$p=0.000\,000\,008<0.05$，因此，建立的回归模型有意义。

2. 多元回归辅助图形命令

残差图命令为 rcoplot，其调用格式为：

```
rcoplot(r,rint)
```

其中，输入参数 r，rint 是多元回归建模命令 regress 输出的结果，运行该命令后展示了残差与置信区间的图形。该命令有助于对建立的模型进行分析，如果图形中出现红色的点，则可以认作异常点，此时可删除异常点，重新建模，最终得到改进的回归模型。

在上面的程序中加入

```
rcoplot(r,rint)
```

得到如图 3-19 所示的图形。

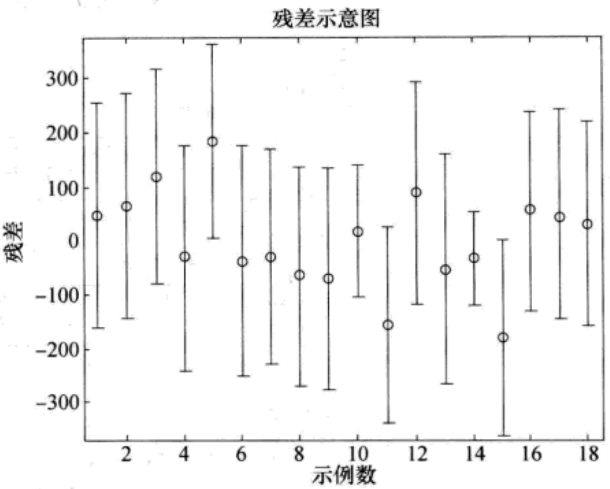


图 3-19 残差与置信区间图

从图 3-19 中可以看到第 5 个点为异常点，实际上从表 3-7 可以发现第 5 个月库存资金额、广告投入、员工薪酬均比 3 月少，为何销售额反而增加？这就可以促使该公司的经理找出原因，寻找对策。下面的例题介绍删除异常点，并对模型进行改进的方法。

例 3.2.2 葛洲坝机组发电耗水率的主要影响因素为库水位、出库流量。现从数据库中将 2005 年 10 月某天 15 时—16 时 06 分范围内的出库流量、库水位对应的耗水率读取出来，见表 3-8，利用多元线性回归分析方法建立耗水率与出库流量、库水位的模型。

表 3-8 耗水率与出库流量、库水位的数据

时间 （年－月－天－时）	库水位 （米）	出库流量 （立方米）	机组发电耗水率 （立方米/万千瓦）
2005－10－某天－15：00	65.08	15 607	60.46
2005－10－某天－15：02	65.10	15 565	60.28
2005－10－某天－15：04	65.12	15 540	60.10
2005－10－某天－15：06	65.17	15 507	59.78
2005－10－某天－15：08	65.21	15 432	59.44
2005－10－某天－15：10	65.37	15 619	59.25
2005－10－某天－15：12	65.38	15 536	58.91
2005－10－某天－15：14	65.39	15 514	58.76
2005－10－某天－15：16	65.40	15 519	58.73
2005－10－某天－15：18	65.43	15 510	58.63
2005－10－某天－15：20	65.47	15 489	58.48
2005－10－某天－15：22	65.53	15 437	58.31
2005－10－某天－16：00	65.62	16 355	57.96
2005－10－某天－16：02	65.58	14 708	57.06
2005－10－某天－16：04	65.70	14 393	56.43
2005－10－某天－16：06	65.84	14 296	55.83

数据来源：余波．多元线性回归分析在机组发电耗水率中的应用．计算机与现代化，2008（2）.

解： 编写程序如下：

```
% 输入原始数据
A=[65.08  15607  60.46
   65.10  15565  60.28
   65.12  15540  60.10
   65.17  15507  59.78
   65.21  15432  59.44
   65.37  15619  59.25
   65.38  15536  58.91
   65.39  15514  58.76
   65.40  15519  58.73
   65.43  15510  58.63
   65.47  15489  58.48
   65.53  15437  58.31
   65.62  16355  57.96
   65.58  14708  57.06
   65.70  14393  56.43
   65.84  14296  55.83];
% 作散点图
subplot(1,2,1),plot(A(:,1),A(:,3),'+ ')
xlabel('x1(库水位)')
ylabel('y(耗水率)')
subplot(1,2,2),plot(A(:,2),A(:,3),'o')
xlabel('x2(出库流量)')
ylabel('y(耗水率)')
```

运行后得到的图形如图 3-20 所示，可以看到无论是库水位还是出库流量都与机组发电耗水率具有线性关系，因此，可以建立机组发电耗水率与库水位和出库流量的二元线性回归模型。

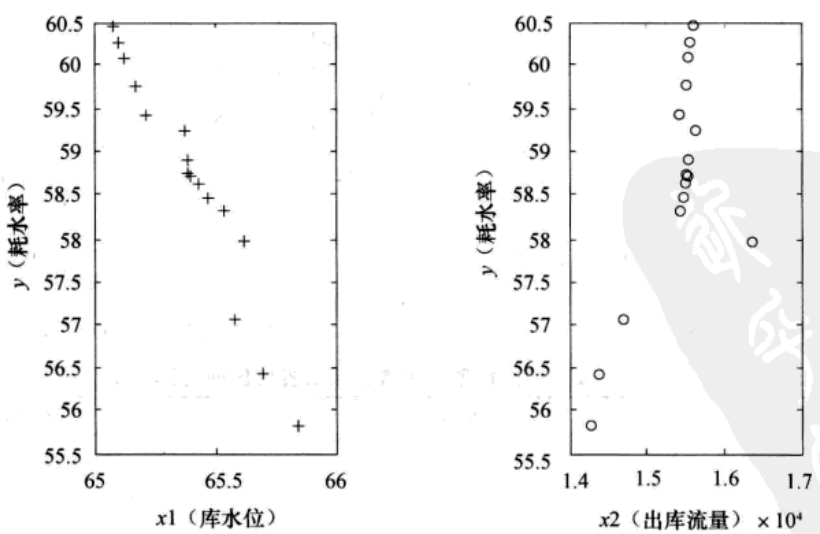


图 3-20 库水位、出库流量与耗水率的散点图

```
% 建立模型
[m,n]= size(A);
y= A(:,3);
x= A(:,1:2);
[b,bint,r,rint,stats]= regress(y,[ones(m,1),x]);
0b,bint,stats
```

输出模型的回归模型的系数、系数置信区间与统计量见表 3-9。

表 3-9　回归模型的系数、系数置信区间与统计量

回 归 系 数	回归系数估计值	回归系数置信区间
β_0	373.869 8	[340.082，407.657 7]
β_1	−4.975 9	[−5.464 2，−4.487 5]
β_2	0.000 7	[0.000 4，0.000 9]
$R^2=0.986\ 3$，$F=468.411\ 8$，$p<0.000\ 1$，$s^2=0.027\ 8$		

由此可得模型为：

$$\hat{y} = 373.869\ 8 - 4.975\ 9x_1 + 0.000\ 7x_2$$

```
% 模型改进
rcoplot(r,rint);
```

得到图形如图 3-21 所示，发现有一个异常点，下面给出删除异常点后，重新建模的程序。

```
% 删除异常点程序并建模
[b1,bint1,r1,rint1,stats1]= regress([y(1:12);y(14:m)],[ones(m- 1,1),[x(1:12,:);x(14:m,:)]])
rcoplot(r1,rint1);
```

删除异常点后，残差示意图如图 3-22 所示，此时没有异常点，改进回归模型的系数、系数置信区间与统计量见表 3-10。

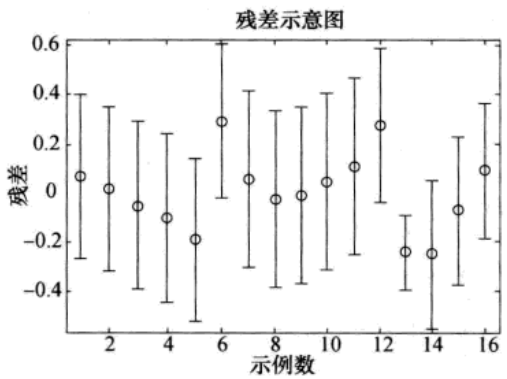


图 3-21　残差示意图

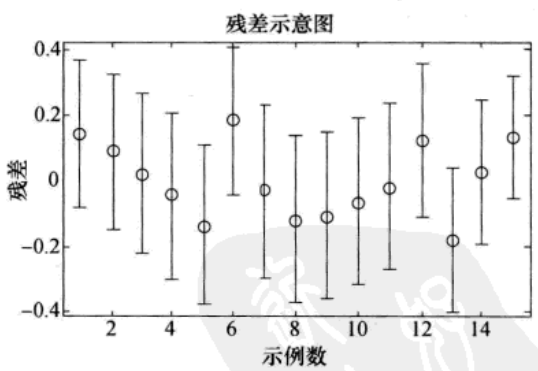


图 3-22　删除异常点后残差示意图

表 3-10　改进回归模型的系数、系数置信区间与统计量

回 归 系 数	回归系数估计值	回归系数置信区间
β_0	328.461 6	[290.614 5，366.308 7]
β_1	−4.359 4	[−4.888 0，−3.830 8]
β_2	0.001 0	[0.000 7，0.001 3]
$R^2=0.993\ 1$，$F=858.584\ 6$，$p<0.000\ 1$，$s^2=0.015\ 0$		

我们将表3-9与表3-10加以比较，可以发现：可决系数从0.986 3提高到0.993 1，F统计量从468.411 8提高到858.584 6，删除异常点后的模型每个参数的置信区间进一步缩小，由此可知改进后的模型显著性提高。

3.2.3 多元线性回归实例

例3.2.3 现代服务业是社会分工不断深化的产物，随着经济的发展，科学技术的进步，现代服务业的发展受到多种因素和条件的影响。不仅受到经济总体发展水平的影响，还受到第二产业、就业、投入等因素的影响，从这几个主要方面出发，利用江苏省统计年鉴的有关数据，通过建立多元线性回归模型对1989—2007年各种因素对现代服务业的影响进行回归分析。假如构建如下江苏省服务业增长模型：

$$Y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\beta_4 x_4 \tag{3.2.4}$$

在(3.2.4)式的模型中，β_0是常数项，β_1、β_2、β_3、β_4表示各种影响因素的常量系数。Y表示江苏省服务业的增加值（单位：亿元），反映了江苏省服务业发展的总体水平。$x_1\sim x_4$表示影响江苏省服务业发展的四种主要因素和影响，其中x_1表示江苏省人均GDP（单位：元），说明江苏省总体经济发展水平对服务业的影响；x_2表示江苏省第二产业的增加值（单位：亿元），主要说明工业发展对服务业的影响，体现生产性服务业的需求规模；x_3表示江苏省服务业的就业人数（单位：万人）；x_4表示江苏省服务业资本形成总额（单位：亿元），主要体现服务业投资的经济效应。

表3-11 江苏省关于服务业发展及各影响因素相关数据

年 份	Y服务业增加值	x_1省人均GDP	x_2第二产业增加值	x_3服务业就业人数	x_4服务业资本形成总额
1989	37.76	2 038	70.24	589.74	252.01
1990	28.13	2 109	35.53	623.19	275.82
1991	93.58	2 353	101.33	640.95	330.71
1992	160.62	3 106	325.34	706.39	439.32
1993	286.58	4 321	478.79	786.37	620.97
1994	277.12	5 801	588.72	855.97	858.91
1995	387.11	7 319	528.49	920.45	1 102.71
1996	367.16	8 471	358.86	975.66	1 293.43
1997	291.77	9 371	337.74	1 025.22	1 370.21
1998	280.01	10 049	228.24	1 102.31	1 624.74
1999	227.61	10 695	280.05	1 151.68	1 773.37
2000	329.16	11 765	515.74	1 192.02	1 903.37
2001	385.44	12 882	471.57	1 263.77	2 131.87
2002	437.02	14 396	697.03	1 341.86	2 189.78
2003	601.39	16 830	1 182.62	1 407.63	2 686.57
2004	704.72	20 223	1 650.88	1 443.37	3 362.19
2005	1 291.11	24 560	1 917.05	1 542.46	3 930.56
2006	1 360.09	28 814	1 895.8	1 625.06	4 628.59
2007	1 769.28	33 928	2 055.56	1 713.33	5 287.91

数据来源：江苏省统计年鉴。

运用MATLAB软件对上述数据进行多元线性回归分析。

解：编写程序如下：

```
% 输入各影响因素的数据
x0= [2038    70.24    589.74    252.01
     2109    35.53    623.19    275.82
     ...
    33928  2055.56   1713.33   5287.91];
y= [37.76,28.13,93.58,160.62,286.58,277.12,387.11,367.16,291.77,280.01,227.61,329.16,385.44,
437.02,601.39,704.72,1291.11,1360.09,1769.28]';        % Y 服务业增加值列向量
[n,p]= size(x0);                                       % 矩阵 x0 的行数即样本容量
x= [ones(n,1),x0];                                     % 构造设计矩阵
[db,dbint,dr,drint,dstats]= regress(y,x);              % 调用多元回归分析命令
```

回归参数的估计

输出：

```
db= 345.2493
     0.1672
     0.1962
   - 0.7012
   - 0.6537
```

即 β 的最小二乘估计为：

$$\begin{aligned}\hat{\beta} &= (\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4)^{\mathrm{T}}\\ &= (345.249,0.167\,2,0.196\,2,-0.701\,2,-0.653\,7)^{\mathrm{T}}\end{aligned}$$

所以，服务业增加值 Y 对 4 个自变量的线性回归方程为

$$\hat{y} = 345.249 + 0.167\,2x_1 + 0.196\,2x_2 - 0.701\,2x_3 - 0.653\,7x_4 \tag{3.2.5}$$

输出：

```
dstats=
        1.0e+ 003 * (0.0010    0.1727    0.0000    5.7926)
```

其中 dstats 的第 4 项是残差的方差估计值。所以，残差方差 σ^2 的无偏估计值为

$$\hat{\sigma}^2 = 5\,792.6$$

下面对例 3.2.3 的回归模型进行显著性检验。接上面的程序，在 MATLAB 命令窗口中继续输入：

```
TSS= y' * (eye(n)- 1/n * ones(n,n)) * y;      % 计算 TSS
H= x * inv((x' * x)) * x';                    % 计算对称幂等矩阵
ESS= y' * (eye(n)- H) * y;                    % 计算 ESS
RSS= y' * (H- 1/n * ones(n,n)) * y;           % 计算 RSS
MSR= RSS/p;                                   % 计算 MSR
MSE= ESS/(n- p- 1);                           % 计算 MSE
% F 检验
F0= (RSS/p)/(ESS/(n- p- 1));                  % 计算 F0
Fa= finv(p,n- p- 1,0.95);                     % F 分布时的临界值 F0.95(p,n-p-1)
% t 检验
S= MSE * inv(x' * x);                         % 计算回归参数的协方差矩阵
T0= db./sqrt(diag(S));                        % 每个回归参数的 T 统计量
Ta= tinv(n- p- 1,0.975);                      % t 分布的分位数
pp= tpdf(T0,n- p- 1);                         % 每个回归参数的 T 统计量对应的概率
```

```
% 可决系数检验
R2= RSS/TSS;                                    % 计算样本可决系数
```

程序的输出结果列在表 3-12 和表 3-13 中。

表 3-12 方差分析表

方差来源	平方和	自由度	均方和	F	p
回归	4 000 513	4	1 000 128.161	172.656	0
误差	81 096.389	14	5 792.599		
总计	4 081 609	18			

表 3-13 回归系数

变量	β	标准差	t	p
常数项	345.25	150.322	2.297	0.038
省人均 GDP	0.167	0.044	3.812	0.002
第二产业增加值	0.196	0.082	2.39	0.031
服务业就业人数	−0.701	0.216	−3.242	0.006
服务业资本形成总额	−0.654	0.295	−2.215	0.044

该方程的拟合优度判定系数

$$R^2 = \frac{SSR}{SST} = 0.98$$

调整后的拟合优度判定系数

$$R_a^2 = 1-(1-R^2)\times\frac{n-1}{n-p-1} = 0.976$$

这说明该多元线性回归方程的拟合程度比较理想。

F 检验：H_0：$\beta_1=\beta_2=\beta_3=\beta_4=0$；$H_1$：$\beta_i$（$i=1$，2，3，4）不全为 0。

从表 3-12 可知，统计量

$$F0 = 172.656$$

给定一个显著性水平 $\alpha=0.05$，查 F 分布表，得到一个临界值 $Fa=3.112\,2$。因为 $F0>Fa$，或者由 $F0$ 的 p 值为 0（小于 0.05），所以拒绝 H_0，接受备择假设。这说明总体回归系数 β_i 不全为 0，即表明模型的线性关系在 95%的置信水平下显著成立。

t 检验：H_0：$\beta_1=0$；H_1：$\beta_1\neq0$。

统计量

$$T0(1) = 2.297$$

给定一个显著性水平 $\alpha=0.05$，查 t 分布表，得到一个临界值 $Ta=2.144\,8$。因为 $T0(1)>Ta$ 或者由 $T0$ 的 p 值为 0.038（小于 0.05），所以拒绝 H_0，接受备择假设，即回归系数 $\beta_1\neq0$。对于其他回归系数 β_i（$i=2$，3，4），用上述同样的方法可以得出各回归系数是显著不为 0 的。

3.3 逐步回归

3.3.1 最优回归方程的选择

在建立经济预测问题的数学模型时，常常从可能影响预测量 Y 的许多因素中挑选一批因

素作为自变量，应用回归分析的方法建立回归方程用于预测或控制。问题是如何在为数众多的因素中挑选变量，以建立所谓的这批观测数据的“最优”回归方程。

什么是“最优”回归方程呢？

从上一节的学习，我们知道，回归方程中所包含的自变量越多，回归平方和（RSS）就越大，剩余平方和（ESS）就越小，一般来讲，剩余均方和（MSE）也随之较小，因而预测就较精确。所以“最优”回归方程中应包括尽可能多的变量，特别是对 Y 有显著影响的变量不能遗漏。但是事情总是一分为二的。方程所含的变量太多，也有不利的一面：第一，在预测时必须测定许多变量，且计算不方便；第二，如果方程中含有对 Y 不起作用或作用极小的变量，那么 ESS 不会由于这些变量的增加而减少多少，相反由于 ESS 自由度的减少，会使 MSE 增大；第三，由于存在着对 Y 不显著的变量，以致影响了回归方程的稳定性，反而使预测效果下降，因此我们又希望在“最优”回归方程中不包含对 Y 影响不显著的变量。

综上所述，所谓“最优”回归方程，就是包含所有对 Y 影响显著的变量而不包含对 Y 影响不显著的变量的回归方程。

选择“最优”回归方程有以下几种方法。

方法 1：从所有可能的变量组合的回归方程中挑选最优者，即把所有包含 1 个，2 个，……，直至所有变量的线性回归方程全部计算出来，对每个方程及自变量作显著性检验，然后从中挑选一个方程，要求该方程中所有的变量全部显著，且 MSE 较小。

这种方法当然可以找到一个“最优”方程，然而计算工作量太大。如果有 10 个因子，就要建立 $2^{10}-1=1023$ 个方程。因此，这种方法只在变量较少时使用。

方法 2：从包含全部变量的回归方程中逐次剔除不显著因子。首先建立包含全部变量的回归方程，然后对每一个因子作显著性检验，剔除不显著因子中偏回归平方和最小的一个因子，重新建立包含全部变量（剔除的除外）的回归方程。然后重复上面的过程，对新建立回归方程的每一个因子作显著性检验，剔除不显著因子中偏回归平方和最小的因子，再重新建立回归方程。如此，当新建立的回归方程中所有因子都显著时，回归方程就是“最优”的了。

这种方法在因子，特别是不显著因子不多时，可以采用。但计算的工作量仍然可能较大。

方法 3：从一个变量开始，把变量逐个引入回归方程。这一方法首先计算各因子与 Y 的相关系数，将绝对值最大的一个因子引入方程，并对回归平方和进行检验，若显著，则引入。然后找出余下的因子中与 Y 的偏相关系数最大的那个因子，将其引入方程，检验显著性，等等，当引入的因子建立的方程检验不显著时，该因子就不再引入。

这种方法尽管工作量较小，但并不保证最后所得到的方程是“最优”的，还得进一步作检验，剔除不显著因子。同时这种方法每一步要计算偏相关系数，也比较麻烦。

结合方法 2 与方法 3，产生了一种建立“最优”回归方程的方法——逐步回归分析。

逐步回归的基本思想是，将变量一个一个引入，引入变量的条件是偏回归平方和经检验是显著的，同时每引入一个新变量后，对已选入的变量要进行逐个检验，将不显著变量剔除。

逐步回归的基本思想是有进有出。具体做法是将变量一个一个引入，每引入一个自变量后，对已引入的变量要进行逐个检验，当原引入的变量由于后面变量的引入而变得不再显著时，要将其剔除。引入一个变量或从回归方程中剔除一个变量为逐步回归的一步，每一步都要进行 F 检验，以确保每次引入新的变量之前回归方程中只包含显著的变量。这个过程反复进行，直到既无显著的自变量引入回归方程，也无不显著的自变量从回归方程中剔除为止。这样就可以保证最后所得的变量子集中的所有变量都是显著的。这样经若干步以后便得“最

优”变量子集。

3.3.2 逐步回归的 MATLAB 方法

逐步回归的计算实施过程可以利用 MATLAB 软件在计算机上自动完成，我们要求关心应用的读者一定要通过前面的叙述掌握逐步回归方法的思想，这样才能用对用好逐步回归法。

在 MATLAB 7.0 统计工具箱中用做逐步回归的命令是 stepwise，它提供了一个交互式画面，通过这个工具你可以自由地选择变量，进行统计分析，其通常用法是：

```
stepwise(X,Y,in,penter,premove)
```

其中 X 是自变量数据，Y 是因变量数据，分别为 $n\times p$ 和 $n\times 1$ 的矩阵，in 是矩阵 X 的列数的指标，给出初始模型中包括的子集，默认设定为全部自变量不在模型中，penter 为变量进入时显著性水平，默认值为 0.05，premove 为变量剔除时显著性水平，默认值为 0.10。

在应用 stepwise 命令进行运算时，程序不断提醒将某个变量加入（move in）回归方程，或者提醒将某个变量从回归方程中剔除（move out）。

注意 *应用 stepwise 命令做逐步回归，数据矩阵 X 的第一列不需要人工加一个全 1 向量，程序会自动求出回归方程的常数项（intercept）。*

下面通过一个例子说明 stepwise 的用法。

例 3.3.1 （Hald，1960）Hald 数据是关于水泥生产的数据。某种水泥在凝固时放出的热量 Y(单位：卡/克）与水泥中 4 种化学成分所占的百分比有关：

X_1：$3CaO\cdot Al_2O_3$

X_2：$3CaO\cdot SiO_2$

X_3：$4CaO\cdot Al_2O_3\cdot Fe_2O_3$

X_4：$2CaO\cdot SiO_2$

在生产中测得 13 组数据，见表 3-14，试建立 Y 关于这些因子的“最优”回归方程。

表 3-14 水泥生产的 Hald 数据

序　号	1	2	3	4	5	6	7	8	9	10	11	12	13
$X1$	7	1	11	11	7	11	3	1	2	21	1	11	10
$X2$	26	29	56	31	52	55	71	31	54	47	40	66	68
$X3$	6	15	8	8	6	9	17	22	18	4	23	9	8
$X4$	60	52	20	47	33	22	6	44	22	26	34	12	12
Y	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

解： 在 MATLAB 软件包中写一个 M 文件“liti3_3_1.m”。

```
X=[7,26,6,60;1,29,15,52;11,56,8,20;11,31,8,47;7,52,6,33;11,55,9,22;3,71,17,6;1,31,22,44;  2,
54,18,22;21,47,4,26;1,40,23,34;11,66,9,12;10,68,8,12];                    % 自变量数据
Y=[78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5,93.1,115.9,83.8,113.3,109.4]';    % 因变量数据
stepwise(X,Y,[1,2,3,4],0.05,0.10)  % in=[1,2,3,4]表示 X1、X2、X3、X4 均保留在模型中
```

程序执行后得到下列逐步回归的窗口，如图 3-23 所示。

在图 3-23 中，用蓝色行显示变量 $X1$、$X2$、$X3$、$X4$ 均保留在模型中，窗口的右侧按钮上方提示：将变量 $X3$ 剔除回归方程（Move $X3$ out），单击 Next Step 按钮，即进行下一步运算，将第 3 列数据对应的变量 $X3$ 剔除回归方程。单击 Next Step 按钮后，剔除的变量 $X3$ 所对应的行用红色表示，同时又得到提示：将变量 $X4$ 剔除回归方程（Move $X4$ out），单击 Next Step 按钮，

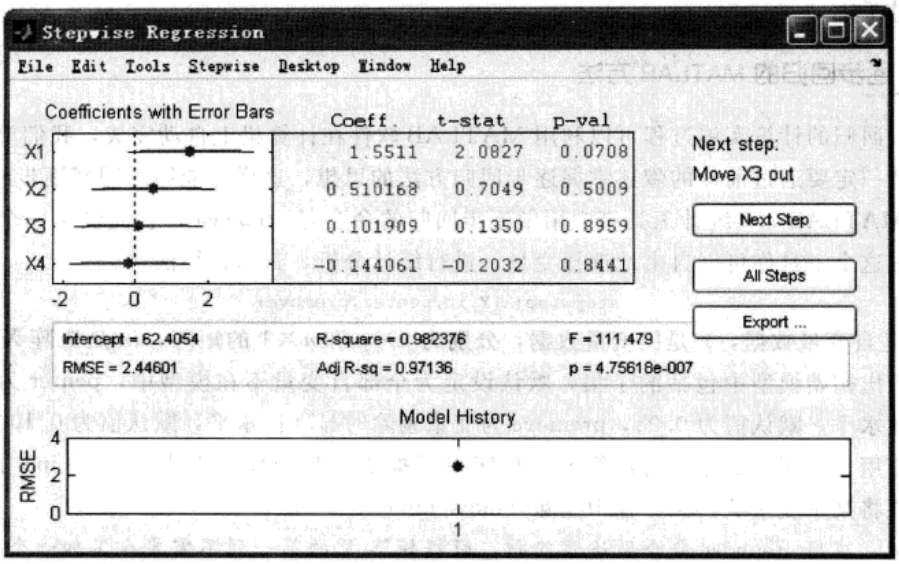


图 3-23　逐步回归窗口

即进行下一步运算，将第 4 列数据对应的变量 $X4$ 剔除回归方程。单击 Next Step 按钮后，$X4$ 所对应的行用红色表示，同时提示：Move no terms，即没有需要加入（也没有需要剔除）的变量了（如图 3-24 所示）。在 MATLAB 7.0 软件包中，可以直接单击 All Steps 按钮，求出结果（省略中间过程）。

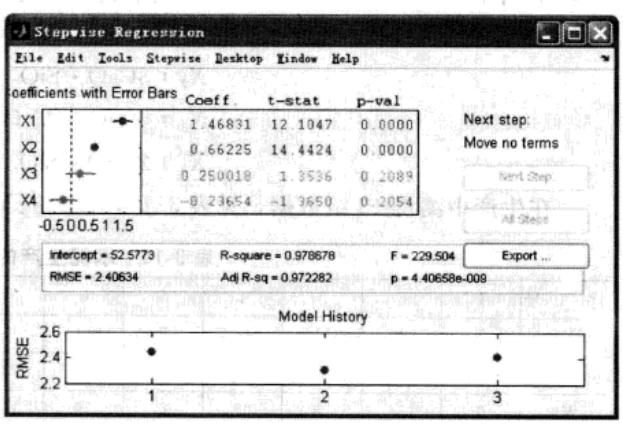


图 3-24　逐步回归结果

由图 3-24，最后得到回归方程（蓝色行是被保留的有效行，红色行表示被剔除的变量）：

$$Y=52.5773+1.46831X_1+0.66225X_2$$

回归方程中录用了原始变量 X_1 和 X_2。

图 3-24 中显示了模型参数分别为：$R^2=0.978\ 678$，修正的 R^2 值 $R_a^2=0.972\ 282$，$F=229.504$，与显著性概率相关的 $p=4.4065\ 8e-009<0.05$，残差均方 RMSE=2.406 34（这个值越小越好）。以上指标值都很好，说明回归效果比较理想。另外，截距 intercept=52.577 3，是回归方程的常数项。

逐步回归窗口中对已建模型给出了在线与超链接的显示功能，当将光标指向“Model History”框中的均方残差 RMSE 的第一个蓝色点时，光标在线显示“in model：$X1$、$X2$、$X3$、$X4$”，若双击光标，则超链接到图 3-22 所示的逐步回归窗口。从“Model History”框中可以观察不同模型的均方残差变化。

习　题　3

1. 以家庭为单位，某种商品年需求量与该商品价格之间的一组调查数据见表 3-15。

表 3-15 某商品年需求量与价格之间的关系

价格 x（元）	5	2	2	2.3	2.5	2.6	2.8	3	3.3	3.5
需求量（kg）	1	3.5	3	2.7	2.4	2.5	2	1.5	1.2	1.2

（1）求经验回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$。

（2）检验线性关系的显著性（$\alpha=0.05$，采用 F 检验）。

2. 某种合金强度与碳含量有关，研究人员在生产试验中收集了该合金的强度 y 与碳含量 x 的数据，见表 3-16。试建立 y 与 x 的函数关系模型，并检验模型的可信度。

表 3-16 合金的强度与碳含量数据表

x	0.10	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.20	0.21	0.23
y	42.0	41.5	45.0	45.5	45.0	47.5	49.0	55.0	50.0	55.0	55.5	60.5

3. 零售商为了解每周的广告费与销售额之间的关系，记录了表 3-17 的统计资料。

表 3-17 每周广告费与销售额统计数据

广告费 X（万）	40	20	25	20	320	50	40	20	50	40	25	50
销售额 Y（百万）	385	400	395	365	475	440	490	420	560	525	480	510

画出散点图，并在 Y 对 X 回归为线性的假定下，用最小二乘法算出一元回归方程。

4. 某省 1978—1989 年消费基金、国民收入使用额和平均人口资料见表 3-18。试配合适当的回归模型并进行各种检验；若 1990 年该省国民收入使用额为 67（十亿元），平均人口为 58（百万人），当显著性水平 $\alpha=0.05$ 时，试估计 1990 年消费基金的预测区间。

表 3-18 某省 1978—1989 年消费基金、国民收入使用额和平均人口资料

年 份	消费基金（十亿元）y	国民收入使用额（十亿元）x_2	平均人口数（百万人）x_3
1978	9.0	12.1	48.20
1979	9.5	12.9	48.90
1980	10.0	16.8	49.54
1981	10.6	14.8	50.25
1982	12.4	16.4	51.02
1983	16.2	20.9	51.84
1984	17.7	24.2	52.76
1985	20.1	28.1	56.39
1986	21.8	30.1	54.55
1987	25.3	35.8	55.35
1988	31.3	48.5	56.16
1989	36.0	54.8	56.98

5. 在生、储、盖、圈、保这五个控制油气聚集条件互相结合可以形成油气藏的条件下，油气藏的储量密度（10^4 T/km^2）与以下生油条件参数有密切关系，这些参数是：生油门限以下平均地温梯度 Δt（用变量 X_1 表示），生油门限以下总有机碳百分含量 C%（用变量 X_2 表示），生油岩体积与沉积岩体积百分比（用变量 X_3 表示），砂泥岩厚度百分比（用变量 X_4 表示），生油门限以下生油带总烃与有机碳的百分比即有机质转化率（用变量 X_5 表示）。根据表 3-19 中的数据，用逐步回归求储量密度 Y 与这五个因素间的回归关系式。

表 3-19 原始数据表

样 品	X_1	X_2	X_3	X_4	X_5	Y
1	3.18	1.15	9.4	17.6	3	0.7
2	3.8	0.79	5.1	30.5	3.8	0.7
3	3.6	1.1	9.2	9.1	3.65	1

（续）

样　品	X_1	X_2	X_3	X_4	X_5	Y
4	2.73	0.73	14.5	12.8	4.68	1.1
5	3.4	1.48	7.6	16.5	4.5	1.5
6	3.2	1	10.8	10.1	8.1	2.6
7	2.6	0.61	7.3	16.1	16.16	2.7
8	4.1	2.3	3.7	17.8	6.7	3.1
9	3.72	1.94	9.9	36.1	4.1	6.1
10	4.1	1.66	8.2	29.4	13	9.6
11	3.35	1.25	7.8	27.8	10.5	10.9
12	3.31	1.81	10.7	9.3	10.9	11.9
13	3.6	1.4	24.6	12.6	12.76	12.7
14	3.5	1.39	21.3	41.1	10	14.7
15	4.75	2.4	26.2	42.5	16.4	21.3

注：原始数据取自我国东部十五个勘探程度相对较高的中、新生代盆地及凹陷。

实验 2　多元线性回归与逐步回归

实验目的

1. 熟练掌握线性回归模型的建立方法，掌握 regress 命令的使用方法。
2. 掌握编程求总离差平方和 TSS、回归平方和 RSS、残差平方和 ESS 等相关统计量。
3. 掌握逐步回归的思想与方法，掌握 stepwise 命令的使用方法。

实验数据与内容

选取 1989—2003 年的全国的统计数据，考虑的自变量包括：（1）工业总产值，设为 $x1$（亿元）；（2）农业总产值，设为 $x2$（亿元）；（3）建筑业总产值，设为 $x3$（亿元）；（4）社会商品零售总额，设为 $x4$（亿元）；（5）全民人口数，设为 $x5$（万人）；（6）受灾面积，设为 $x6$（万公顷）；（7）国家财政收入，设为 y（亿元）。数据见表 3-20，（1）建立多元回归模型；（2）用逐步回归求国家财政收入 y 与 6 个因素间的回归关系式。

表 3-20　1989—2003 年统计数据

年　份	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	y
1989	6 484.00	4 100.60	794.00	8 101.40	112 704.0	46 991.00	2 664.90
1990	6 858.00	4 954.30	859.40	8 300.10	114 333.0	38 474.00	2 937.10
1991	8 087.10	5 146.40	1 015.10	9 415.60	115 823.0	55 472.00	3 149.48
1992	10 284.50	5 588.00	1 415.00	10 993.70	117 171.0	51 333.00	3 483.37
1993	14 143.80	6 605.10	2 284.70	12 462.10	118 517.0	48 829.00	4 348.95
1994	19 359.60	9 169.20	3 012.60	16 264.70	119 850.0	55 043.00	5 218.10
1995	24 718.30	11 884.60	3 819.60	20 620.00	121 121.0	45 821.00	6 242.20
1996	29 082.60	13 539.80	4 530.50	24 774.10	122 389.0	46 989.00	7 407.99
1997	32 412.10	13 852.50	4 810.60	27 298.90	123 626.0	53 429.00	8 651.14
1998	33 387.90	14 241.90	5 231.40	29 152.50	124 761.0	50 145.00	9 875.95
1999	35 087.20	14 106.20	5 470.60	31 134.70	125 786.0	49 981.00	11 444.08
2000	39 047.30	13 873.60	5 888.00	34 152.60	126 743.0	54 688.00	13 395.23
2001	42 374.60	14 462.80	6 375.40	37 595.20	127 627.0	52 215.00	16 386.04
2002	45 975.20	14 931.50	7 005.00	42 027.10	128 453.0	47 119.00	18 903.64
2003	53 092.90	14 870.10	8 181.30	45 842.00	129 227.0	54 506.00	21 715.25

数据来源：http://www.stats.gov.cn/tjsj/ndsj/2009/indexch.htm.

第4章 判别分析

同回归分析一样，判别分析也是一种重要的统计分析方法。这一方法的基本思想是根据已知类别的样本所提供的信息，总结出分类的规律性，建立判别公式和判别准则，判别新的样本点所属类型。本章介绍距离判别分析、贝叶斯（Bayes）判别分析及它们在 MATLAB 软件中的实现。

4.1 距离判别分析

4.1.1 判别分析的概念

在一些自然科学和社会科学的研究中，研究对象用某种方法已划分为若干类型。得到的一个新样品数据（通常是多元的）后，要确定该样品属于已知类型中的哪一类，这样的问题属于判别分析。

在生产、科研和日常生活中经常需要根据观测到的数据资料，对所研究的对象进行分类。例如，在经济学中，根据人均国民收入、人均工农业产值、人均消费水平等多种指标来判定一个国家的经济发展程度所属类型；在地质勘探中，根据岩石标本的多种特性来判别地层的地质年代，由采样分析出的多种成分来判别此地是有矿还是无矿，是铜矿还是铁矿等；在油田开发中，根据钻井的电测或化验数据，判别是否遇到油层、水层、干层或油水混合层；在农林害虫预报中，根据以往的虫情、多种气象因子来判别一个月后的虫情是大发生、中发生还是正常；在体育运动中，判别某游泳队的“苗子”是适合练蛙泳、仰泳还是自由泳等；在医疗诊断中，根据某人多种体检指标（如体温、血压、白血球等）来判别此人是有病还是无病。总之，在实际问题中需要判别的问题几乎到处可见。

从统计数据分析的角度，以上问题可概括为如下模型：设有 k 个总体 $G_1,G_2,\cdots,G_k$，它们都是 p 维总体，其数量指标为

$$X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$$

设 X 在各个总体下具有不同的分布特征，一般说来，各个总体 G_i 的分布是未知的，需要由各个总体取得的样本（训练样本）来估计其均值向量与协方差矩阵，对于某一新样品数据 $x=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$，要根据各总体的分布特征按一定判别准则判断它来自哪一个总体。对各个总体的特征进行分析与建立判别准则的过程是判别分析的主要内容，我们将在下面作全面介绍。

判别分析按判别的组数来区分，有两组判别和多组判别；按不同总体所用的数学模型来区分，有线性判别和非线性判别；按判别时处理变量的方法不同，有逐步判别和序贯判别等。判别分析可以从不同角度提出问题，因此有不同判别准则，如马氏距离最小准则、Fisher 准

则、平均损失最小准则、最小平方准则、最大似然准则等。本章仅介绍两种常用的判别方法：距离判别法和贝叶斯判别法。

4.1.2 距离的定义

1. 闵可夫斯基距离 (Minkowski distance)

设有 n 维向量 $x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，$y=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，称

$$d_1(x,y)=\sum_{i=1}^{n}|x_i-y_i|$$

为 n 维向量 x、y 之间的绝对距离 (absolute distance)。

称

$$d_2(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

为 n 维向量 x、y 之间的欧氏距离 (Euclidean distance)。

称

$$d_r(x,y)=\left(\sum_{i=1}^{n}|x_i-y_i|^r\right)^{1/r}$$

为 n 维向量 x、y 之间的闵可夫斯基距离 (简称闵氏距离)，其中 $r(r>0)$ 为常数。

显然，当 $r=2$ 和 $r=1$ 时闵可夫斯基距离分别为欧氏距离和绝对距离。

2. 马氏距离 (Mahalanobis distance)

马氏距离是由印度统计学家普拉山塔·钱德拉·马哈拉诺比斯 (Prasantha Chandra Mahalanobis) 提出的，由于马氏距离具有统计意义，在距离判别分析时马氏距离比欧氏距离应用得更多。以下给出三种情形的马氏距离定义。

(1) 同一总体的两个向量之间的马氏距离

设总体 G 的两个 n 维观测向量 $x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，$y=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，称

$$d(x,y)=\sqrt{(x-y)\Sigma^{-1}(x-y)^{\mathrm{T}}} \tag{4.1.1}$$

为 n 维向量 x、y 之间的马氏距离。其中，Σ 为总体 G 的协方差矩阵，Σ^{-1} 为 Σ 的逆矩阵。

在 (4.1.1) 式中，Σ 为实对称正定矩阵。当 Σ 为单位矩阵时，马氏距离就是欧氏距离。

(2) 一个向量到一个总体的马氏距离

设 x 是取自均值向量为 μ、协方差矩阵为 Σ 的总体 G 的 n 维观测向量，则称

$$d(x,G)=\sqrt{(x-\mu)\Sigma^{-1}(x-\mu)^{\mathrm{T}}} \tag{4.1.2}$$

为 n 维向量 x 与总体 G 的马氏距离。

(3) 两个总体之间的马氏距离

设两个总体 G_1、G_2 的均值向量分别为 μ_1、μ_2，协方差矩阵相等且皆为 Σ，则总体 G_1、G_2 之间的马氏距离定义为

$$d(G_1,G_2)=\sqrt{(\mu_1-\mu_2)\Sigma^{-1}(\mu_1-\mu_2)^{\mathrm{T}}} \tag{4.1.3}$$

显然，在 (4.1.2) 式和 (4.1.3) 式中，当 Σ 为单位矩阵时，马氏距离就化为通常的欧氏距离。

在 MATLAB 中，计算马氏距离的命令为 mahal，其调用格式为：

```
d= mahal(Y,X)
```

该命令计算 X 矩阵中每一个点 (行) 至 Y 矩阵中每一个点 (行) 的马氏距离。其中 Y 的

列数必须等于 X 的列数，但它们的行数可以不同。X 的行数必须大于列数。输出 d 是距离向量。

需要注意的是，在进行判别分析时，一般不采用欧氏距离，其原因在于该距离与量纲有关。例如，平面上有 A、B、C、D 四个点，横坐标代表重量（单位：kg），纵坐标代表长度（单位：cm），如图 4-1 所示。

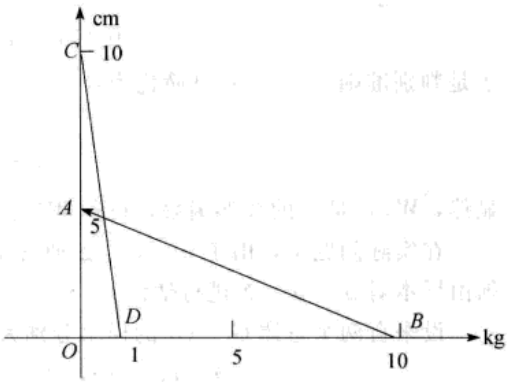


图 4-1 欧氏距离

这时，A 点与 B 点的距离 AB，C 点与 D 点的距离 CD 分别为

$$AB = \sqrt{5^2+10^2} = \sqrt{125}$$

$$CD = \sqrt{10^2+1^2} = \sqrt{101}$$

显然 $AB>CD$。

现在，如果长度以 mm 为单位，重量的单位保持不变，于是 A 点的坐标为（0，50），C 点的坐标为（0，100），此时计算线段 AB 与 CD 的“长度”分别为

$$AB = \sqrt{50^2+10^2} = \sqrt{2600}$$

$$CD = \sqrt{100^2+1^2} = \sqrt{10\,001}$$

显然 $AB<CD$。

两种不同的度量单位，得到的距离大小关系相反，这表明欧氏距离是有缺陷的，即当向量的分量是不同的量纲时，欧氏距离的大小与指标的单位有关。可以证明马氏距离与量纲无关，因此在实践中常常选用马氏距离进行判别分析。

4.1.3 两总体的距离判别分析

距离判别分析的基本思想是：首先根据已知分类的数据，分别计算各类的重心即分组（类）的均值，其次对任给的一次观测，计算其与每一类重心的距离，最后依据最小距离进行判别。若它与第 i 类的距离最小，就认为它来自第 i 类。

设 G_1、G_2 为两个不同的 p 维总体，且 G_i 的均值向量是 μ_i，协方差矩阵是 $\Sigma_i(i=1,2)$。$x=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$ 是一个待判样品，按马氏距离定义判别准则为

$$\begin{cases} x \in G_1 & \text{若 } d(x,G_1) \leqslant d(x,G_2) \\ x \in G_2 & \text{若 } d(x,G_2) < d(x,G_1) \end{cases} \tag{4.1.4}$$

即当 x 到 G_1 的马氏距离不超过到 G_2 的马氏距离时，判 x 来自 G_1；反之，判 x 来自 G_2。

由于马氏距离与总体的协方差矩阵有关，所以利用马氏距离进行判别分析需要分别考虑两个总体的协方差矩阵是否相等。

1. 两个总体协方差矩阵相等的情况

设两个总体 G_1、G_2 的协方差矩阵均为 Σ，由（4.1.2）式，考虑样品 x 到两个总体的马氏距离平方差

$$\begin{aligned} d^2(x,G_2)-d^2(x,G_1) &= (x-\mu_2)^{\mathrm{T}}\Sigma^{-1}(x-\mu_2)-(x-\mu_1)^{\mathrm{T}}\Sigma^{-1}(x-\mu_1) \\ &= 2\left[x-\frac{1}{2}(\mu_1+\mu_2)\right]^{\mathrm{T}}\Sigma^{-1}(\mu_1-\mu_2) \end{aligned}$$

$$= 2(x-\bar{\mu})^{\mathrm{T}}\Sigma^{-1}(\mu_1-\mu_2)$$

其中 $\bar{\mu}=\frac{1}{2}(\mu_1+\mu_2)$。令

$$W(x) = (x-\bar{\mu})^{\mathrm{T}}\Sigma^{-1}(\mu_1-\mu_2) \tag{4.1.5}$$

于是判别准则（4.1.4）可简化为：

$$\begin{cases} x \in G_1 & W(x) \geqslant 0 \\ x \in G_2 & W(x) < 0 \end{cases} \tag{4.1.6}$$

显然，$W(x)$是 x 的线性函数，也称 $W(x)$为线性判别函数。

在实际问题中，由于 μ_1、μ_2、Σ 通常是未知的，数据资料来自两个总体的样本，因此，必须由样本对 μ_1、μ_2、Σ 进行估计。

设来自两个总体 G_1、G_2 的样本分别为

$$x_1^{(1)}, x_2^{(1)}, \cdots, x_{n_1}^{(1)} \in G_1, \quad x_1^{(2)}, x_2^{(2)}, \cdots, x_{n_2}^{(2)} \in G_2$$

于是两个样本的均值向量与协方差矩阵分别为

$$\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i^{(1)} = \bar{x}^{(1)}, \hat{\mu}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_i^{(2)} = \bar{x}^{(2)} \tag{4.1.7}$$

$$S_1 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i^{(1)}-\bar{x}^{(1)})(x_i^{(1)}-\bar{x}^{(1)})^{\mathrm{T}}$$

$$S_2 = \frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_i^{(2)}-\bar{x}^{(2)})(x_i^{(2)}-\bar{x}^{(2)})^{\mathrm{T}} \tag{4.1.8}$$

当 $\Sigma_1=\Sigma_2=\Sigma$ 时，Σ 的一个无偏估计记为 S，则

$$\Sigma = \hat{S} = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2} \tag{4.1.9}$$

称 S 为混合样本方差。

这样，线性判别函数 $W(x)$的估计为

$$\hat{W}(x) = \left(x-\frac{1}{2}(\bar{x}^{(1)}+\bar{x}^{(2)})\right)^{\mathrm{T}} S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)}) \tag{4.1.10}$$

于是，两个总体的距离判别准则（4.1.6）化为

$$\begin{cases} x \in G_1 & \hat{W}(x) \geqslant 0 \\ x \in G_2 & \hat{W}(x) < 0 \end{cases} \tag{4.1.11}$$

在 MATLAB 中，进行数据的判别分析命令为 classify，其调用格式为：

```
class = classify(sample,training,group 'type')
```

将 sample 数据的每一行指定到训练集 training 的一个类中。sample 和 training 必须具有相同的列数。group 向量包含从 1 到组数的正整数，它指明训练集中的每一行属于哪一个类。group 和 training 必须具有相同的行数。'type'是可选项，type 选'linear'表示总体为多元正态总体（默认是这一选项），type 选'quadratic'与'mahalanobis'分别表示 FitsMVN densities with covariance estimates stratified by group 与 Uses Mahalanobis distances with stratified covariance estimates。该函数返回 class，它是一个与 sample 具有相同行数的向量。class 的每一个元素指定 sample 中对应元素的分类。通过计算 sample 与 training 中每一行的马氏距离，classify 函数决定 sample 中的每一行属于哪一个分类。

以下举例说明应用（4.1.11）式进行判别的 MATLAB 程序的编写。

例 4.1.1 （1989 年国际数学竞赛 A 题：蠓的分类）蠓是一种昆虫，分为很多类型，其中有一种名为 Af，是能传播花粉的益虫；另一种名为 Apf，是会传播疾病的害虫。这两种类型

的蠓在形态上十分相似，很难区别。现测得6只Apf和9只Af蠓虫的触角长度和翅膀长度数据。Apf：(1.14，1.78)、(1.18，1.96)、(1.20，1.86)、(1.26，2.00)、(1.28，2.00)、(1.30，1.96)；Af：(1.24，1.72)、(1.36，1.74)、(1.38，1.64)、(1.38，1.82)、(1.38，1.90)、(1.40，1.70)、(1.48，1.82)、(1.54，1.82)、(1.56，2.08)。

若两类蠓虫的协方差矩阵相等，试判别以下的三只蠓虫属于哪一类。

(1.24，1.8)、(1.28，1.84)、(1.4，2.04)

解：(方法一) 按照两总体的距离判别法，直接编写程序。假定两类蠓虫是两个总体，且两个总体的协方差矩阵相等，根据判别准则（4.1.11）式，源程序如下：

```
apf= [1.14,1.78;1.18,1.96;1.20,1.86;1.26,2.00;1.28,2.00;1.30,1.96];        % 总体 apf
af= [1.24,1.72;1.36,1.74;1.38,1.64;1.38,1.82;1.38,1.90;1.40,1.70;1.48,1.82;1.54,1.82;1.56,2.
08];                                        % 总体 af
x= [1.24,1.8;1.28,1.84;1.4,2.04];           % 输入原始待判数据
n1= size(apf,1);                            % 总体 apf 的样本容量
n2= size(af,1);                             % 总体 af 的样本容量
m1= mean(apf);                              % 总体 apf 的均值向量
m2= mean(af);                               % 总体 af 的均值向量
s1= cov(apf);                               % 总体 apf 的协方差
s2= cov(af);                                % 总体 af 的协方差
s= ((n1- 1) * s1+ (n2- 1) * s2)/(n1+ n2- 2); % 计算样本均值与协方差矩阵
for i= 1:3
     W(i)= (x(i,:)- 1/2 * (m1+ m2)) * inv(s) * (m1- m2)';     % 计算判别函数值
end
```

输出结果为：

```
W=
   2.1640    1.3568    1.9802
```

判别函数值 W 的每一项都大于0，由判别准则（4.1.11）可知，三只蠓虫均属于Apf。

(方法二) 直接调用MATLAB的判别分析命令classify。程序如下：

```
apf= [1.14,1.78;1.18,1.96;1.20,1.86;1.26,2.00;1.28,2.00;1.30,1.96];        % 总体 apf
af= [1.24,1.72;1.36,1.74;1.38,1.64;1.38,1.82;1.38,1.90;1.40,1.70;1.48,1.82;1.54,1.82;1.56,
2.08];                                      % 总体 af
training= [apf;af];                         % 合并两个总体形成训练集
n1= size(apf,1);                            % 总体 apf 中样本的行数
n2= size(af,1);                             % 总体 af 中样本的行数
group= [ones(1,n1), 2 * ones(1,n2)];        % apf 中样本与 af 中样本类属
x= [1.24,1.8;1.28,1.84; 1.4,2.04];          % 输入原始待判数据即 sample
class = classify(x,training,group)          % 判别分析
```

输出结果为：

```
class =
        1
        1
        1
```

表明三只蠓虫均属于Apf。

2. 两个总体协方差矩阵不相等的情况

设 $\Sigma_1 \neq \Sigma_2$，由（4.1.2）式，样品 x 到两个总体 G_1、G_2 的马氏距离平方分别为：

$$d_1^2(x) = d^2(x,G_1) = (x-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(x-\mu_1)$$
$$d_2^2(x) = d^2(x,G_2) = (x-\mu_2)^{\mathrm{T}}\Sigma_2^{-1}(x-\mu_2)$$

令

$$\begin{aligned} W(x) &= d^2(x,G_2) - d^2(x,G_1) \\ &= (x-\mu_2)^{\mathrm{T}}\Sigma_2^{-1}(x-\mu_2) - (x-\mu_1)^{\mathrm{T}}\Sigma_1^{-1}(x-\mu_1) \end{aligned} \tag{4.1.12}$$

此时，判别准则（4.1.11）可化为

$$\begin{cases} x \in G_1 & W(x) \geqslant 0 \\ x \in G_2 & W(x) < 0 \end{cases} \tag{4.1.13}$$

由于 $W(x)$是 x 的二次函数，故称为二次判别函数。

在实际问题中，由于 μ_1、μ_2、Σ_1、Σ_2 通常是未知的，可用各总体的训练样本做估计，即分别以 $\bar{x}^{(1)}$、$\bar{x}^{(2)}$估计 μ_1、μ_2，以 S_1、S_2 估计 Σ_1、Σ_2，利用式（4.1.7）～式（4.1.9），得到 $d_1^2(x)$、$d_2^2(x)$和判别函数 $W(x)$的估计分别为：

$$\hat{d}_1^2(x) = (x-\bar{x}^{(1)})^{\mathrm{T}}S_1^{-1}(x-\bar{x}^{(1)}) \tag{4.1.14}$$

$$\hat{d}_2^2(x) = (x-\bar{x}^{(2)})^{\mathrm{T}}S_2^{-1}(x-\bar{x}^{(2)}) \tag{4.1.15}$$

$$\hat{W}(x) = (x-\bar{x}^{(2)})^{\mathrm{T}}S_2^{-1}(x-\bar{x}^{(2)}) - (x-\bar{x}^{(1)})^{\mathrm{T}}S_1^{-1}(x-\bar{x}^{(1)}) \tag{4.1.16}$$

则判别准则（4.1.13）为

$$\begin{cases} x \in G_1 & \hat{W}(x) \geqslant 0 \\ x \in G_2 & \hat{W}(x) < 0 \end{cases} \tag{4.1.17}$$

例 4.1.2 假定两类总体的协方差矩阵不相等，重新判别例 4.1.1 中三只蠓虫的类别。

解： 根据题意假设，按照（4.1.17）的判别准则，源程序如下：

```
apf= [1.14,1.78; 1.18,1.96;1.20,1.86;1.26,2.00;1.28,2.00;1.30,1.96];
af= [1.24,1.72;1.36,1.74;1.38,1.64;1.38,1.82;1.38,1.90;1.40,1.70;1.48,1.82;1.54,1.82;1.56,
2.08];
x= [1.24,1.8;1.28,1.84;1.4,2.04];                    % 输入原始数据
W= mahal(x,apf)- mahal(x,af)                          % 计算判别函数值
```

输出结果为：

```
W =
    1.7611
    3.8812
    3.6468
```

由判别准则（4.1.17）可知，三只蠓虫均属于 Af。

3. 两个总体协方差矩阵相等的检验

例 4.1.1 和例 4.1.2 的结果大相径庭，由此我们不禁要问究竟哪个结果是合理的。问题的关键在于：两类蠓虫总体的协方差矩阵是否相等？下面介绍协方差矩阵相等的检验方法，用各 p 元总体的样本做估计，即 S_1、S_2 估计 Σ_1、Σ_2，检验的假设为

$$\text{原假设 } H_0: S_i = S; \text{备择假设 } H_1: S_i \neq S (i = 1,2) \tag{4.1.18}$$

其中 $S = \dfrac{(n_1-1)\ S_1 + (n_2-1)\ S_2}{n_1+n_2-2}$。

检验统计量

$$Q_i = (n_i - 1)[\ln|S| - \ln|S_i| - p + \mathrm{tr}(S^{-1}S_i)](i = 1,2) \tag{4.1.19}$$

可以证明，当 n_i 较大时，Q_i 分布服从 $\chi^2(p(p+1)/2)$。对给定的显著性水平 α，若 $Q_i < \chi^2_{1-\alpha}(p(p+1)/2)(i=1,2)$，则接受 H_0；否则拒绝 H_0。考虑到这里的分布是近似分布，在两样本容量均不小于5时使用上述检验是适当的。

对于例4.1.1的数据，检验协方差矩阵是否相等的源程序如下：

```
n1= 6;n2= 9;p= 2;
s= ((n1- 1) * s1+ (n2- 1) * s2)/(n1+ n2- 2);                    % 计算混合样本方差
Q1= (n1- 1) * (log(det(s))- log(det(s1))- p+ trace(inv(s) * s1)) ;
Q2= (n2- 1) * (log(det(s))- log(det(s2))- p+ trace(inv(s) * s2)) ;
                                                                % 计算检验统计量观测值
```

输出结果为：

```
Q1=
    2.5784
Q2=
    0.7418
```

给定 $\alpha=0.05$，查表得到临界值 $\chi^2_{1-\alpha}(3)=7.8147$（命令 chi2inv（0.95，3））。

由于 $Q_1<7.8147$，$Q_2<7.8147$，故认为两类总体协方差矩阵相同。显然，例4.1.1的解法更为合理些。

因此，在进行距离判别过程中，首先要检验各总体的协方差矩阵是否相等，从而确定是采用线性判别函数还是二次判别函数。

4.1.4 多个总体的距离判别分析

设有 k 个总体 $G_1,G_2,\cdots,G_k$，均值向量分别为 $\mu_1,\mu_2,\cdots,\mu_k$，协方差矩阵分别为 $\Sigma_1,\Sigma_2,\cdots,\Sigma_k$。对于待判别的样品 x，计算其到各总体的马氏距离，若存在第 m 个总体使得

$$d(x,G_m) = \min_{1\leqslant i\leqslant k} d(x,G_i) \tag{4.1.20}$$

则判别样品 x 属于第 m 个总体。类似于两个总体的判别，对于多个总体的协方差矩阵也应首先检验是否相等，具体检验方法参考第2章的2.2.2节。

1. 总体协方差矩阵相等时的判别

设有 k 个总体 $G_1,G_2,\cdots,G_k$，均值向量分别为 $\mu_1,\mu_2,\cdots,\mu_k$，协方差矩阵分别为 $\Sigma_1,\Sigma_2,\cdots,\Sigma_k$。若 $\Sigma_1=\Sigma_2=\cdots=\Sigma_k=\Sigma$，由式（4.1.5）可知新样品 x 到 G_j 和 G_i 的马氏距离的平方差为：

$$d^2(x,G_j) - d^2(x,G_i) = 2\left[\left(x - \frac{1}{2}(\mu_i+\mu_j)\right)\right]^{\mathrm{T}}\Sigma^{-1}(\mu_i - \mu_j)$$

令判别函数

$$W_{ij}(x) = \left[(x - \frac{1}{2}(\mu_i+\mu_j)\right]^{\mathrm{T}}\Sigma^{-1}(\mu_i - \mu_j) \tag{4.1.21}$$

则 x 到 G_i 的距离最小等价于对所有的 $j(j\neq i)$，有 $W_{ij}(x)>0$。

由于总体的均值向量与协方差矩阵是未知的，所以用样本的均值和协方差矩阵代替。设 $x_1^{(j)},x_2^{(j)},\cdots,x_{n_j}^{(j)}$ 是取自总体 $G_j(j=1,2,\cdots,k)$ 的训练样本，记

$$\bar{x}^{(j)} = \frac{1}{n_j}\sum_{j=1}^{n_j} x_i^{(j)}, j = 1,2,\cdots,k \tag{4.1.22}$$

$$S_j=\frac{1}{n_j-1}\sum_{j=1}^{n_i}(x_i^{(j)}-\overline{x}^{(j)})(x_i^{(j)}-\overline{x}^{(j)})^{\mathrm{T}}\quad(j=1,2,\cdots,k)\tag{4.1.23}$$

$$S=\sum_{j=1}^{k}(n_j-1)S_j/(n-k)\tag{4.1.24}$$

于是得到判别函数 $W_{ij}(x)$ 的估计为

$$\hat{W}_{ij}(x)=[(x-\frac{1}{2}(\overline{x}^{(i)}+\overline{x}^{(j)})]^{\mathrm{T}}S^{-1}(\hat{\overline{x}}^{(i)}-\hat{\overline{x}}^{(j)})\tag{4.1.25}$$

判别准则为：对所有的 $j(j\neq i)$，$\hat{W}_{i,j}(x)>0$，则判别 $x\in G_i$。

例 4.1.3 依据例 2.2.2 表 2-6 中给出的身体指标化验数据，对三个待判数（190，67，30，17），（315，100，35，19），（240，60，37，18）进行判别归类。

解： 根据例 2.2.2 的结论，可以认为三类总体协方差矩阵相等。程序如下：

```
A=[260   75   40   18   310   122   30   21   320   64   39   17;
   200   72   34   17   310    60   35   18   260   59   37   11;
   ...
   260  135   39   29   280    40   37   17   250  117   36   16];
G1= A(:,1:4);G2= A(:,5:8);G3= A(:,9:12);                    % 三类总体数据
x=[190 67 30 17;315 100 35 19;240 60 37 18];                % 待判定的数据
m(1,:)= mean(G1);m(2,:)= mean(G2);m(3,:)= mean(G3);
s1= cov(G1);s2= cov(G2);s3= cov(G3);              % 计算样本均值与协方差矩阵
s= 19*(s1+ s2+ s3)/57;                            % 计算混合样本方差
for i= 1:3
    for j= 1:3
        for k= 1:3
            w(j,k)= (x(i,:)- 1/2*(m(j,:)+ m(k,:)))*inv(s)*(m(j,:)- m(k,:))'; % 计算判别
                函数
            if w(j,k)< 0
                q= 0;break;
            else q= 1;
            end
        end
        if q== 1
        y(i)= j;
        end
    end
end
y
```

输出结果为：

```
y =
    1    3    2
```

由以上判别准则可知，三个待判数据（190，67，30，17），（315，100，35，19），（240，60，37，18）分别属于 G_1，G_3 和 G_2。

2. 总体协方差矩阵不全相等时的判别

此时，样品 x 到各个总体 G_i（$i=1,2,\cdots,k$）的马氏距离平方分别为：

$$d^2(x,G_i)=(x-\mu_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(x-\mu_i) \tag{4.1.26}$$

判别准则为：$\min\limits_{1\leqslant i\leqslant k}d(x,\ G_i)=d(x,\ G_m)$，则判别 $x\in G_m$。

在实际情况中，用训练样本对总体作估计。得到二次判别函数 $d^2(x,\ G_i)$的估计：

$$d^2(x,G_i)=(x-\bar{x}^{(i)})^{\mathrm{T}}S^{-1}(x-\bar{x}^{(i)})\quad(i=1,2,\cdots,k)$$

若

$$d^2(x,G_j)=\min_{1\leqslant i\leqslant k}d^2(x,G_i) \tag{4.1.27}$$

则判别 $x\in G_j$。

4.2 判别准则的评价

当一个判别准则提出以后，还要研究它的优良性，即考察它的误判率。以训练样本为基础的误判率的估计思想：若属于 G_1 的样品被误判为属于 G_2 的个数为 N_1 个，属于 G_2 的样品被误判为属于 G_1 的个数为 N_2 个，两类总体的样品总数为 N，则误判率 p 的估计为：

$$\hat{p}=\frac{N_1+N_2}{N}$$

针对具体情况，通常采用回代法和交叉法进行误判率的估计。

1. 回代误判率

设 G_1，G_2 为两个总体，$x_1,x_2,\cdots,x_m$ 和 $y_1,y_2,\cdots,y_n$ 是分别来自 G_1，G_2 的训练样本，以全体训练样本作为 $m+n$ 个新样品，逐个代入已建立的判别准则中判别其归属，这个过程称为回判。回判结果中若属于 G_1 的样品被误判为属于 G_2 的个数为 N_1 个，属于 G_2 的样品被误判为属于 G_1 的个数为 N_2 个，则误判率估计为：

$$\hat{p}=\frac{N_1+N_2}{m+n} \tag{4.2.1}$$

误判率的回代估计易于计算。但是，$\hat{p}$是由建立判别函数的数据反过来用作评估准则的数据而得到的。因此$\hat{p}$作为真实误判率的估计是有偏的，往往比真实误判率小。当训练样本容量较大时，$\hat{p}$可以作为真实误判率的一种估计，具有一定的参考价值。

2. 交叉误判率

交叉误判率估计是每次剔除一个样品，利用其余的 $m+n-1$ 个训练样本建立判别准则，再用所建立的准则对删除的样品进行判别。对训练样本中每个样品都做如上分析，以其误判的比例作为误判率，具体步骤如下：

1）从总体 G_1 的训练样本开始，剔除其中一个样品，剩余的 $m-1$ 个样品与 G_2 中的全部样品建立判别函数；

2）用建立的判别函数对剔除的样品进行判别；

3）重复步骤 1）和 2），直到 G_1 中的全部样品依次被删除又进行判别，其误判的样品个数记为 N_1^*；

4）对 G_2 的样品重复步骤 1）、2）和 3），直到 G_2 中的全部样品依次被删除又进行判别，其误判的样品个数记为 N_2^*。

于是交叉误判率估计为

$$\hat{p}^*=\frac{N_1^*+N_2^*}{m+n} \tag{4.2.2}$$

用交叉法估计真实误判率是较为合理的。

例 4.2.1 根据表 4-1 的数据，判别两类总体的协方差矩阵是否相等，然后用马氏距离判别未知地区的类别，并计算回代误判率与交叉误判率。

表 4-1 各地区农、林、牧、渔各业数据

类 别	农	林	牧	渔	类 别	农	林	牧	渔
2	89.70	9.50	105.20	9.60	1	405.90	11.30	236.40	5.80
2	86.70	1.50	60.80	20.60	1	450.60	15.70	224.60	20.10
2	95.50	3.50	88.40	40.10	1	529.50	73.70	195.90	308.80
2	191.30	12.30	96.30	1.70	1	688.00	66.20	371.60	132.30
2	307.60	26.10	216.20	6.00	1	433.20	82.30	215.50	330.50
2	141.30	43.30	58.20	82.30	1	405.90	54.00	226.10	104.30
2	250.40	11.20	154.40	15.20	1	658.30	27.10	352.60	134.80
2	337.40	23.60	114.10	3.80	1	665.70	51.90	480.30	85.20
2	254.00	8.60	80.90	1.10	1	817.90	56.80	423.20	390.10
2	28.90	1.80	32.50	0.10	1	439.90	39.40	292.30	101.20
2	49.40	3.50	30.30	2.10	1	769.90	50.90	605.00	41.00
2	348.80	10.10	134.00	3.90	x	431.30	47.20	210.60	14.40
2	899.40	34.00	685.90	61.20	x	1 401.30	47.20	654.70	350.70
2	1 142.70	30.80	448.50	334.20	x	1 331.60	57.00	693.80	20.40
1	503.10	21.80	332.30	188.50	x	279.90	15.10	118.50	5.10

解：首先判断两组数据协方差是否相等；再建立判别准则，计算回代和交叉误判率，源程序如下：

```
a=[503.10  21.80  332.30  188.50
   405.90  11.30  236.40    5.80
   ...
   769.90  50.90  605.00   41.00];
b=[89.70   9.50  105.20    9.60
   86.70   1.50   60.80   20.60
   ...
  1142.70  30.80  448.50  334.20];
x=[431.30  47.20  210.60   14.40
  1401.30  47.20  654.70  350.70
  1331.60  57.00  693.80   20.40
   279.90  15.10  118.50    5.10];          % 输入两类总体数据及待判数据
n1=length(a(:,1));
n2=length(b(:,1));                          % 样本容量
s1=cov(a);
s2=cov(b);                                  % 样本协方差
p=4;
s=((n1-1)*s1+(n2-1)*s2)/(n1+n2-2);          % 混合样本协方差
q1=(n1-1)*(log(det(s))-log(det(s1))-p+trace(inv(s)*s1));
q2=(n2-1)*(log(det(s))-log(det(s2))-p+trace(inv(s)*s2));  % 检验统计量
chi2inv(0.95,10);                                         % 验证两总体的协方差矩阵相同
for i=1:4
    D(i)=(x(i,:)-mean(a))*inv(s)*(x(i,:)-mean(a))'-(x(i,:)-mean(b))*inv(s)*(x
```

```
      (i,:)- mean(b))';
end
D
```

输出判别函数值

```
D =
    - 1.2313   - 4.7511   - 4.8792    4.0777
```

由 D 结果可得：$x1$、$x2$、$x3$ 属于第一类，$x4$ 属于第二类。

```
%%%%%%%%%%计算回代误判率,源程序如下%%%%%%%%%%
for i= 1:n1
d11(i)= (a(i,:)- mean(a)) * inv(s) * (a(i,:)- mean(a))'- (a(i,:)- mean(b)) * inv(s) * (a(i,:)-
        mean(b))';
end
for i= 1:n2
        d22(i)= (b(i,:)- mean(b)) * inv(s) * (b(i,:)- mean(b))'- (b(i,:)- mean(a)) * inv(s) *
              (b(i,:)- mean(a))';
end
n11= length(find(d11> 0));n22= length(find(d22> 0));
p0= (n11+ n22)/(n1+ n2)
%%%%%%%%%%计算交叉误判率,源程序如下%%%%%%%%%%
for i= 1:n1
       A= a([1:i- 1,i+ 1:n1],:);
       n1= length(A(:,1));n2= length(b(:,1));
       s1= cov(A);s2= cov(b);p= 4;
       s= ((n1- 1) * s1+ (n2- 1) * s2)/(n1+ n2- 2);
       D11(i)= (a(i,:)- mean(A)) * inv(s) * (a(i,:)- mean(A))'- (a(i,:)- mean(b)) * inv(s) * (a
              (i,:)- mean(b))';
end
for i= 1:n2
       B= b([1:i- 1,i+ 1:n2],:);
       n1= length(a(:,1));n2= length(B(:,1));
       s1= cov(A);s2= cov(B);p= 4;
       s= ((n1- 1) * s1+ (n2- 1) * s2)/(n1+ n2- 2);
       % 计算判别函数
       D22(i)= (b(i,:)- mean(B)) * inv(s) * (b(i,:)- mean(B))'- (b(i,:)- mean(a)) * inv(s) * (b
              (i,:)- mean(a))';
end
N11= length(find(D11> 0));N22= length(find(D22> 0));
p1= (N11+ N22)/(n1+ n2)
```

输出结果：

```
p0 =
    0.1923
p1 =
    0.2400
```

在本例中，回代误判率是 0.192 30，交叉误判率是 0.240 0。由此可见，该判别准则是有效的。

4.3　贝叶斯判别分析

距离判别只要求知道总体的数字特征，不涉及总体的分布函数。当参数和协方差未知时，

就用样本的均值和协方差矩阵来估计。距离判别方法简单实用，但它没有考虑每个总体出现的机会大小（先验概率），也没有考虑到错判的损失。贝叶斯判别正是为了解决这两个问题提出的判别方法。下面，我们先介绍两总体的贝叶斯判别法。

4.3.1 两总体的贝叶斯判别

1. 任意分布总体的一般讨论

考虑两个 p 元总体 G_1、G_2 分别具有概率密度函数 $f_1(x)$、$f_2(x)$，又设 G_1、G_2 出现的先验概率为

$$p_1 = P(G_1), \quad p_2 = P(G_2)$$

其中 $p_1+p_2=1$。

当取得新样品 $x=(x_1,x_2,\cdots,x_p)^{\mathrm{T}}$ 后，根据贝叶斯公式，G_1、G_2 的后验概率分别为

$$P(G_1 \mid x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}, \quad P(G_2 \mid x) = \frac{p_2 f_2(x)}{p_1 f_1(x) + p_2 f_2(x)} \tag{4.3.1}$$

因此，两个总体的贝叶斯判别准则为

$$\begin{cases} x \in G_1 : \text{若 } P(G_1 \mid x) \geqslant P(G_2 \mid x) \\ x \in G_2 : \text{若 } P(G_1 \mid x) < P(G_2 \mid x) \end{cases} \tag{4.3.2}$$

2. 两个正态总体的贝叶斯判别

（1）两个总体协方差矩阵相等的情形

设总体 G_1、G_2 的协方差矩阵相等且为 Σ，概率密度函数为

$$f_j(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_j)^{\mathrm{T}}\Sigma^{-1}(x-\mu_j)\right\} \quad (j=1,2)$$

总体 G_1、G_2 的先验概率 $p_1=P(G_1)$，$p_2=P(G_2)$（$p_1+p_2=1$），则基于两正态总体误判损失相等的贝叶斯判别准则为

$$\begin{cases} x \in G_1, & \text{当 } w_1(x) \geqslant w_2(x) \\ x \in G_2, & \text{当 } w_1(x) < w_2(x) \end{cases} \tag{4.3.3}$$

其中 $w_j(x)=(\overline{x}^{(j)})^{\mathrm{T}}\Sigma^{-1}x-\frac{1}{2}(\overline{x}^{(j)})^{\mathrm{T}}\Sigma^{-1}\overline{x}^{(j)}+\ln p_j (j=1,2)$。

而基于两正态总体后验概率的贝叶斯判别准则为

$$\begin{cases} x \in G_1, & \text{当 } d_1^2(x) \leqslant d_2^2(x) \\ x \in G_2, & \text{当 } d_1^2(x) > d_2^2(x) \end{cases} \tag{4.3.4}$$

其中 $d_j^2(x)=(x-\overline{x}^{(j)})^{\mathrm{T}}\Sigma^{-1}(x-\overline{x}^{(j)})-2\ln p_j (j=1,2)$。

在实际问题中，关于先验概率 p_1、p_2，通常用下列两种方式选取：

1）采用等概率选取，即 $p_1=p_2=\frac{1}{2}$。

2）按训练样本的容量 $n_1,n_2,\cdots n_k$ 的比例选取，即

$$p_1 = \frac{n_1}{n_1+n_2}, \quad p_2 = \frac{n_2}{n_1+n_2}$$

由于 μ_1、μ_2、Σ 通常是未知的，可用各总体的训练样本均值 $\overline{x}^{(1)}$、$\overline{x}^{(2)}$ 估计 μ_1、μ_2，混合样本方差 S 估计 Σ。

例 4.3.1 重新对例 4.1.1 的三只蠓虫的类别进行贝叶斯判别（假设误判损失相等）。

解： 第 1 步，可以验证两个总体服从二元正态分布（第 2 章的正态性检验，读者自证）；

第 2 步，检验两个总体的协方差矩阵相等（见例 4.1.3）；

第 3 步，估计两个总体的先验概率$\hat{p}_1$、$\hat{p}_2$，这里按样本容量的比例选取。由于 Apf 与 Af 分别为 6 只与 9 只，故估计 Apf 类蠓虫的先验概率$\hat{p}_1=\frac{6}{6+9}=0.4$，Af 类蠓虫的先验概率$\hat{p}_2=\frac{9}{6+9}=0.6$；

第 4 步，利用 MATLAB 软件计算。程序如下：

```
% 输入原始数据
apf= [1.14,1.78; 1.18,1.96;1.20,1.86;1.26,2.00;1.28,2.00;1.30,1.96];
af= [1.24,1.72;1.36,1.74;1.38,1.64;1.38,1.82;1.38,1.90;1.40,1.70;1.48,1.82;1.54,1.82;1.56,
2.08];
x= [1.24,1.8;1.28,1.84; 1.4,2.04];                          % 输入待判数据
n1= size(apf,1);
n2= size(af,1);
p1= n1/(n1+ n2);
p2= n2/(n1+ n2);                                            % 先验概率
m1= mean(apf);m2= mean(af);                                 % 样本均值向量
s1= cov(apf);s2= cov(af);                                   % 样本协方差
s= ((n1- 1) * s1+ (n2- 1) * s2)/(n1+ n2- 2);                % 计算混合样本方差
for i= 1:3
      w1(i)= m1 * inv(s) * x(i,:)'- 1/2 * m1 * inv(s) * m1'+ log(p1);
      w2(i)= m2 * inv(s) * x(i,:)'- 1/2 * m2 * inv(s) * m2'+ log(p2); % 基于误判损失相等的判别函数
    if w1(i)> = w2(i)
        disp(['第',num2str(i),'只蠓虫属于 Apf 类']);
    else
        disp(['第',num2str(i),'只蠓虫属于 Af 类']);
    end
end
```

输出结果：

```
第 1 只蠓虫属于 Apf 类
第 2 只蠓虫属于 Apf 类
第 3 只蠓虫属于 Apf 类
```

（2）两个总体协方差矩阵不相等的情形

设总体 G_1、G_2 的协方差矩阵不相等分别为 Σ_1、Σ_2，概率密度函数为：

$$f_j(x)=\frac{1}{(2\pi)^{p/2}|\Sigma_j|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(x-\mu_j)\right\}\quad(j=1,2)$$

则基于两正态总体误判损失相等的贝叶斯判别准则

$$\begin{cases}x\in G_1:\text{若 } d_1^2(x)\leqslant d_2^2(x)\\ x\in G_2:\text{若 } d_1^2(x)>d_2^2(x)\end{cases}\tag{4.3.5}$$

其中 $d_j^2(x)=(x-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(x-\mu_j)-\ln|\Sigma_j|-2\ln p_j(j=1,2)$。

下面举例说明贝叶斯判别分析的应用。

例 4.3.2　对破产的企业收集它们在破产前两年的年度财务数据，对财务良好的企业也收集同一时间的数据。数据涉及四个变量：$X_1=CF/TD$（现金流量/总债务）、$X_2=NI/TA$（净收益/总资产）、$X_3=CA/CL$（流动资产/流动债务）以及 $X_4=CA/NS$（流动资产/净销售额），数据见表 4-2。假定两总体 G1、G2 均服从四元正态分布，在误判损失相等且先验概率按比例

分配的条件下，对待判样本进行贝叶斯判别。

表 4-2 两类企业财务状况数据

G1（破产企业）				G2（非破产企业）				待判			
X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
−0.45	−0.41	1.09	0.45	0.51	0.10	2.49	0.54	−0.23	−0.30	0.33	0.18
−0.56	−0.31	1.51	0.16	0.08	0.02	2.01	0.53	0.15	0.05	2.17	0.55
0.06	0.02	1.01	0.40	0.38	0.11	3.27	0.35	−0.28	−0.23	1.19	0.66
−0.07	−0.09	1.45	0.26	0.19	0.05	2.25	0.33	0.48	0.09	1.24	0.18
−0.10	−0.09	1.56	0.67	0.32	0.07	4.24	0.63				
−0.14	−0.07	0.71	0.28	0.12	0.05	2.52	0.69				
0.04	0.01	1.50	0.71	−0.02	0.02	2.05	0.35				
−0.06	−0.06	1.37	0.40	0.22	0.08	2.35	0.40				
−0.13	−0.14	1.42	0.44	0.17	0.07	1.80	0.52				

解：第 1 步，检验两个总体的协方差矩阵是否相等。

程序如下：

```
A= [- 0.45  - 0.41  1.09  0.45  0.51  0.10  2.49  0.54
    - 0.56  - 0.31  1.51  0.16  0.08  0.02  2.01  0.53
    ...
    - 0.13  - 0.14  1.42  0.44  0.17  0.07  1.80  0.52]          % 样本数据
x= [- 0.23  - 0.30  0.33  0.18
      0.15    0.05  2.17  0.55
    - 0.28  - 0.23  1.19  0.66
      0.48    0.09  1.24  0.18];                  %输入原始数据和待判数据
G1= A(:,1:4);G2= A(:,5:8);                       %输入二类总体数据
m1= mean(G1);                                    %总体 G1 的均值向量
m2= mean(G2);                                    %总体 G2 的均值向量
s1= cov(G1);                                     %总体 G1 的协方差
s2= cov(G2);                                     %总体 G2 的协方差
n1= size(G1,1);                                  %总体 G1 的样本数
n2= size(G2,1);                                  %总体 G2 的样本数
n= n1+ n2;                                       %两个总体合并的样本数
p= 2;

s= ((n1- 1) * s1+ (n2- 1) * s2)/(n1+ n2- 2);                    % 计算混合样本方差
Q1= (n1- 1) * (log(det(s))- log(det(s1))- p+ trace(inv(s) * s1)) ;
Q2= (n2- 1) * (log(det(s))- log(det(s2))- p+ trace(inv(s) * s2)) ; % 计算检验统计量观测值
% 判断协方差是否相等
if Q1< chi2inv(0.95,p * (p+ 1)/2)&&Q2< chi2inv(0.95,p * (p+ 1)/2)
      disp('两组数据协方差相等');
else
      disp('两组数据协方差不全相等');
end;
```

输出结果：

```
两组数据协方差不全相等
```

第 2 步，根据第 1 步结论，构造判别函数，得出判别结果。程序如下：

```
p1= n1/n;p2= n2/n;                                          % 计算先验概率
for i= 1:4
    d1(i)= mahal(x(i,:),G1)- log(det(s1))- 2 * log(p1);
    d2(i)= mahal(x(i,:),G2)- log(det(s2))- 2 * log(p2);     % 计算判别函数
    if d1(i)< = d2(i)
        disp(['第',num2str(i),'个属于破产企业']);
    else
        disp(['第',num2str(i),'个属于非破产企业']);
  end
end
```

输出结果：

```
第 1 个属于破产企业
第 2 个属于非破产企业
第 3 个属于破产企业
第 4 个属于非破产企业
```

4.3.2　多个总体的贝叶斯判别

1. 一般讨论

设 p 维总体 $G_1, G_2, \cdots, G_k$，G_j 的概率密度为 $f_j(x)$，$j=1,2,\cdots,k$。各总体出现的先验概率为

$$p_j = P(G_j) \quad (j=1,2,\cdots,k)$$

满足 $\sum_{j=1}^{k} p_j = 1$。又由贝叶斯公式，当出现样品 x 时，总体 G_i 的后验概率

$$P(G_i \mid x) = \frac{p_i f_i(x)}{\sum_{j=1}^{k} p_j f_j(x)}$$

此时判定 x 来自后验概率最大的那个总体 G_i，这符合贝叶斯统计推断原则，即贝叶斯判别准则为：若

$$P(G_i \mid x) = \max_{1 \leqslant j \leqslant k} \{P(G_j \mid x)\} \quad (i=1,2,\cdots,k) \tag{4.3.6}$$

则判样本 $x \in G_i$。

注意　当达到最大后验概率的 G_i 不止一个时，可判为达到最大后验概率的总体的任何一个。

2. 多个正态总体的贝叶斯判别

1）当 $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$ 时，设 $G_j \sim N_p(\mu_j, \Sigma)$（$j=1,2,\cdots,k$）。线性判别函数为

$$W_j(x) = a_j^{\mathrm{T}} x + b_j$$

其中 $a_j^{\mathrm{T}} = \mu_j^{\mathrm{T}} \Sigma^{-1}$，$b_j = -\frac{1}{2} \mu_j^{\mathrm{T}} \Sigma^{-1} \mu_j + \ln p_j$（$j=1,2,\cdots,k$）

基于误判损失相等的贝叶斯判别准则为

$$x \in G_i, \text{若 } W_i(x) = \max_{1 \leqslant j \leqslant k} \{W_j(x)\} \tag{4.3.7}$$

基于后验概率的贝叶斯判别准则为

$$x \in G_i, \text{若 } d_i^2(x) = \min_{1 \leqslant j \leqslant k} \{d_j^2(x)\} \tag{4.3.8}$$

其中 $d_j^2(x) = (x-\mu_j)^{\mathrm{T}} \Sigma^{-1} (x-\mu_j) - 2\ln p_j$（$j=1,2,\cdots,k$）。

在实际问题中，由于 $\mu_1,\mu_2,\cdots,\mu_k$ 及 Σ 未知，可用各总体的训练样本均值 $\bar{x}^{(1)},\bar{x}^{(2)},\cdots\bar{x}^{(k)}$ 及 S 估计。

2）当 $\Sigma_1,\Sigma_2,\cdots\Sigma_k$ 不全相等时，设 $G_j\sim N_p(\mu_j,\Sigma_j)$（$j=1,2,\cdots,k$），则基于后验概率的贝叶斯判别准则为

$$x\in G_i,\text{若 } d_i^2(x)=\min_{1\leqslant j\leqslant k}\{d_j^2(x)\} \tag{4.3.9}$$

其中 $d_j^2(x)=(x-\mu_j)^{\mathrm{T}}\Sigma_j^{-1}(x-\mu_j)+\ln|\Sigma_j|-2\ln p_j$。

例 4.3.3 某医院利用心电图检测来对人群进行划分，数据见表 4-3。其中 g 指标表示的是对人群组别的划分，“$g=1$”表示健康人，“$g=2$”表示主动脉硬化患者，“$g=3$”表示冠心病患者，其余两个指标 X_1，X_2 表示测得的心电图中表明心脏功能的两项不相关的指标。某受试者心电图该两项指标的数据分别为 380.20，9.08。设先验概率按比例分配，进行贝叶斯判别，判定其归属。

表 4-3 24 人心电图数据

编　号	X_1	X_2	g	编　号	X_1	X_2	g
1	261.01	7.36	1	13	258.69	7.16	2
2	185.39	5.99	1	14	355.54	9.43	2
3	249.58	6.11	1	15	476.69	11.32	2
4	137.13	4.35	1	16	316.12	8.17	2
5	231.34	8.79	1	17	274.57	9.67	2
6	231.38	8.53	1	18	409.42	10.49	2
7	260.25	10.02	1	19	330.34	9.61	3
8	259.51	9.79	1	20	331.47	13.72	3
9	273.84	8.79	1	21	352.50	11.00	3
10	303.59	8.53	1	22	347.31	11.19	3
11	231.03	6.15	1	23	189.59	5.46	3
12	308.90	8.49	2	24	380.20	9.08	待判

解：编写程序如下：

```
A=[261.01  7.36
   185.39  5.99
   …
   189.59  5.46];
x=[380.20  9.08];                    % 待判样品数据
G1=A(1:11,:);G2=A(12:18,:);G3=A(19:23,:);          % 输入三类总体数据
n1=size(G1,1);              % 总体 G1 的样本数
n2=size(G2,1);              % 总体 G2 的样本数
n3=size(G3,1);              % 总体 G3 的样本数
n=n1+n2+n3;                 % 三个总体合并的样本数
k=3;                        % 总体个数
p=2;
f=p*(p+1)*(k-1)/2;
d=(2*p^2+3*p-1)*(1/(n1-1)+1/(n2-1)+1/(n3-1)-1/(n-k))/(6*(p+1)*(k-1));
p1=n1/n;p2=n2/n;p3=n3/n;
m1=mean(G1);m2=mean(G2);m3=mean(G3);
```

```
s1= cov(G1);s2= cov(G2);s3= cov(G3);                    % 计算协方差阵
s= ((n1- 1) * s1+ (n2- 1) * s2+ (n3- 1) * s3)/(n- k);
M= (n- k) * log(det(s))- ((n1- 1) * log(det(s1))+ (n2- 1) * log(det(s2))+ (n3- 1) * log(det
   (s3)));
T= (1- d) * M                                           % 计算统计量观测值
C= chi2inv(0.95,f)
if  T< chi2inv(0.95,f)
      disp('三组数据协方差相等');
else
      disp('三组数据协方差不全相等');
end;
w(1)= m1 * inv(s) * x'- 1/2 * m1 * inv(s) * m1'+ log(p1);
w(2)= m2 * inv(s) * x'- 1/2 * m2 * inv(s) * m2'+ log(p2);
w(3)= m3 * inv(s) * x'- 1/2 * m3 * inv(s) * m3'+ log(p3);  % 计算判别函数
for i= 1:3
      if w(i)= = max(w)
      disp(['待判样品属于第',num2str(i),'组']);
      end;
end;
```

输出结果：

```
三组数据协方差相等
待判样品属于第 2 组
```

4.3.3　平均误判率

贝叶斯判别的有效性可以通过平均误判率来确定。这里仅对两个正态总体 G_1、G_2，且协方差矩阵相等的情况下研究平均误判率的计算。

设总体 $G_i \sim N_p(\mu_j, \Sigma)(i=1, 2)$，总体 G_1、G_2 的先验概率 $p_1=P(G_1)$，$p_2=P(G_2)(p_1+p_2=1)$，两个总体 G_1、G_2 的马氏平方距离记为

$$\delta = (\mu_1 - \mu_2)^{\mathrm{T}}\Sigma^{-1}(\mu_1 - \mu_2) \tag{4.3.10}$$

则基于误判损失相等时的平均误判率为

$$\begin{aligned} p^* &= P(2|1)p_1 + P(1|2)p_2 \\ &= p_1\Phi\left(\frac{d-\frac{\delta}{2}}{\sqrt{\delta}}\right) + p_2\left[1-\Phi\left(\frac{d+\frac{\delta}{2}}{\sqrt{\delta}}\right)\right] \end{aligned} \tag{4.3.11}$$

其中 $d=\ln p_1 - \ln p_2$，$\Phi(\cdot)$ 为标准正态分布函数。

μ_1、μ_2、Σ 通常是未知的，分别以 $\bar{x}^{(1)}$、$\bar{x}^{(2)}$、S 估计 μ_1、μ_2、Σ，得到 δ 的估计

$$\hat{\delta} = (\bar{x}^{(1)} - \bar{x}^{(2)})^{\mathrm{T}}S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$$

这样，以 $\hat{\delta}$ 替代 δ，计算平均误判概率。

需要指出的是，从（4.3.11）式知，当总体 G_1、G_2 的马氏平方距离 δ 越大，即两总体的分离程度越大时，平均误判概率越小。推广到一般情况也成立。因此，判别准则的误判率在一定程度上依赖于所考虑的各总体间的差异程度。各总体间差异越大，就越有可能建立有效的判别准则。如果各总体间差异很小，做判别分析的意义不大。

例 4.3.4　2008 年全国部分地区城镇居民人均年家庭收入情况见表 4-4。按四种指标分为

两类，用贝叶斯判别判定青海、广东两省区属于哪一类，并用回代法和交叉法对误判率进行估计（假定误判损失相等）。

表 4-4 2008 年全国各省、区、市城镇居民人均年家庭收入 （单位：元/人）

地 区	工薪收入	经营净收入	财产性收入	转移性收入	类 型
北 京	18 738.96	778.36	452.75	7 707.87	1
上 海	21 791.11	1 399.14	369.12	6 199.77	1
天 津	12 849.73	863.52	256.87	7 203.93	2
江 苏	12 319.86	1 999.61	307.31	5 548.78	2
浙 江	15 538.83	3 161.87	1 324.94	4 955.14	2
福 建	12 668.82	2 185.13	952.91	3 879.29	2
山 东	12 940.62	1 194.40	346.90	3 067.05	2
西 藏	12 314.69	303.34	138.08	891.42	2
河 北	8 891.50	1 078.67	224.86	3 946.39	3
山 西	9 019.35	983.21	202.31	3 654.11	3
内蒙古	10 284.43	1 555.31	324.64	3 031.05	3
辽 宁	9 494.59	1 483.30	248.04	4 610.32	3
黑龙江	7 393.39	1 241.37	122.83	3 506.48	3
安 徽	9 302.38	959.43	293.92	3 603.72	3
江 西	9 105.96	1 106.31	265.35	2 985.96	3
河 南	9 043.52	1 161.96	156.46	3 545.86	3
湖 北	9 474.81	1 114.68	244.13	3 340.65	3
湖 南	9 070.97	1 575.08	316.48	3 614.74	3
重 庆	10 957.62	788.26	205.94	3 265.92	3
宁 夏	8 793.54	1 856.94	182.67	3 285.49	3
广 西	10 321.20	1 314.40	441.15	3 316.44	3
四 川	9 117.00	1 040.14	262.90	3 265.06	3
贵 州	7 811.16	770.86	110.90	3 492.70	3
云 南	8 596.88	1 165.96	849.45	3 505.74	3
陕 西	9 794.82	544.00	151.46	3 356.85	3
甘 肃	8 354.63	638.76	65.33	2 610.61	3
新 疆	9 422.22	938.15	141.75	1 976.49	3
青 海	8 595.48	763.07	50.17	3 458.63	待判
广 东	15 188.39	2 405.92	701.25	3 382.95	待判

解：第 1 步，检验三个总体的协方差矩阵是否相等。程序如下：

```
A=[18738.96   778.36  452.75  7707.87
   21791.11  1399.14  369.12  6199.77
   ...
    9422.22   938.15  141.75  1976.49];
x=[8793.54  1856.94  182.67  3285.49
   15188.39  2405.92  701.25  3382.95];          % 待判样品
G1= A(1:2,:);G2= A(3:8,:);G3= A(9:27,:);        % 输入三类总体数据
    n1= size(G1,1);                  % 总体 G1 的样本数
    n2= size(G2,1);                  % 总体 G2 的样本数
```

```
    n3= size(G3,1);                    % 总体 G3 的样本数
    n= n1+ n2+ n3;                     % 三个总体合并的样本数
k= 3;
p= 4;
f= p*(p+1)*(k-1)/2;
d= (2*p^2+3*p-1)*(1/(n1-1)+1/(n2-1)+1/(n3-1)-1/(n-k))/(6*(p+1)*(k-1));
p1= n1/n;p2= n2/n;p3= n3/n;
m1= mean(G1);m2= mean(G2);m3= mean(G3);
s1= cov(G1);s2= cov(G2);s3= cov(G3);                    % 计算协方差矩阵
s= ((n1-1)*s1+(n2-1)*s2+(n3-1)*s3)/(n-k);
M= (n-k)*log(det(s))-((n1-1)*log(det(s1))+(n2-1)*log(det(s2))+(n3-1)*log(det
   (s3)));
T= (1-d)*M                                              % 计算统计量观测值
C= chi2inv(0.95,f)
if T< chi2inv(0.95,f)
     disp('三组数据协方差矩阵相等');
else
     disp('三组数据协方差矩阵不全相等');
end
```

输出结果：

```
三组数据协方差矩阵相等
```

第 2 步，根据第 1 步结论，构造判别函数，得出判别结果。程序如下：

```
for i= 1:2
            w(1)= m1*inv(s)*x(i,:)'-1/2*m1*inv(s)*m1'+log(p1);
            w(2)= m2*inv(s)*x(i,:)'-1/2*m2*inv(s)*m2'+log(p2);
            w(3)= m3*inv(s)*x(i,:)'-1/2*m3*inv(s)*m3'+log(p3);        % 计算判别函数
    for j= 1:3
          if w(j)= = max(w)
          disp(['待判样品属于第',num2str(j),'类城市']);
          end
     end
end
```

输出结果：

```
待判样品属于第 3 类城市
待判样品属于第 2 类城市
```

第 3 步，计算回代误判率。程序如下：

```
n11= 0; n22= 0;n33= 0;
for i= 1:n1
       w1(i,1)= m1*inv(s)*G1(i,:)'-1/2*m1*inv(s)*m1'+log(p1);
       w1(i,2)= m2*inv(s)*G1(i,:)'-1/2*m2*inv(s)*m2'+log(p2);
       w1(i,3)= m3*inv(s)*G1(i,:)'-1/2*m3*inv(s)*m3'+log(p3);    % 计算判别函数
       for j= 1:3
              if  w1(i,j)= = max(w1(i,:))&j~= 1
                   n11= n11+1;
              end
```

```
        end
end
for i= 1:n2
        w2(i,1)= m1 * inv(s) * G2(i,:)'- 1/2 * m1 * inv(s) * m1'+ log(p1);
        w2(i,2)= m2 * inv(s) * G2(i,:)'- 1/2 * m2 * inv(s) * m2'+ log(p2);
        w2(i,3)= m3 * inv(s) * G2(i,:)'- 1/2 * m3 * inv(s) * m3'+ log(p3);    % 计算判别函数
        for j= 1:3
              if  w2(i,j)= = max(w2(i,:))&j~ = 2
                  n22= n22+ 1;
              end
        end
end
for i= 1:n3
        w3(i,1)= m1 * inv(s) * G3(i,:)'- 1/2 * m1 * inv(s) * m1'+ log(p1);
        w3(i,2)= m2 * inv(s) * G3(i,:)'- 1/2 * m2 * inv(s) * m2'+ log(p2);
        w3(i,3)= m3 * inv(s) * G3(i,:)'- 1/2 * m3 * inv(s) * m3'+ log(p3);    % 计算判别函数
        for j= 1:3
              if  w3(i,j)= = max(w3(i,:))&j~ = 3
                  n33= n33+ 1;
              end
        end
end
p00= (n11+ n22+ n33)/(n1+ n2+ n3)
```

输出结果：

```
p00 =
        0
```

第 4 步，计算交叉误判率。程序如下：

```
N11= 0; N22= 0;N33= 0;
for k= 1:n1
        A= G1([1:k- 1,k+ 1:n1],:);
        N1= length(A(:,1));
        M1= mean(A,1);s11= cov(A);
        S1= ((N1- 1) * s11+ (n2- 1) *s2+ (n3- 1) *s3)/(N1+ n2+ n3- k);
        P01= N1/(n- 1);P02= n2/(n- 1);P03= n3/(n- 1);                    % 计算先验概率
        for i= 1:n1
                W1(i,1)= M1 *inv(S1) *G1(i,:)'- 1/2 *M1 *inv(S1) *M1'+ log(P01);
                W1(i,2)= m2 *inv(S1) *G1(i,:)'- 1/2 *m2 *inv(S1) *m2'+ log(P02);
                W1(i,3)= m3 *inv(S1) *G1(i,:)'- 1/2 *m3 *inv(S1) *m3'+ log(P03); % 计算判别函数
                for j= 1:3
                        if  W1(i,j)== max(W1(i,:))&j~ = 1
                            N11= N11+ 1;
                        end
                end
          end
end
```

```
for k= 1:n2
        B= G2([1:k- 1,k+ 1:n2],:);
        N2= length(B(:,1));
        M2= mean(B,1);s22= cov(B);
        S2= ((n1- 1) * s1+ (N2- 1) * s22+ (n3- 1) * s3)/(n1+ N2+ n3- k);  % 计算混合样本方差
        P01= n1/(n- 1);P02= N2/(n- 1);P03= n3/(n- 1);                     % 计算先验概率
        for i= 1:n2
                W2(i,1)= m1 *inv(S2) *G2(i,:)'- 1/2 *m1 *inv(S2) *m1'+ log(P01);
                W2(i,2)= M2 *inv(S2) *G2(i,:)'- 1/2 *M2 *inv(S2) *M2'+ log(P02);
                W2(i,3)= m3 *inv(S2) *G2(i,:)'- 1/2 *m3 *inv(S2) *m3'+ log(P03); % 计算判别函数
                for j= 1:3
                        if  W2(i,j)= = max(W2(i,:))&j~ = 2
                            N22= N22+ 1;
                        end
                end
        end
end
for k= 1:n3
        C= G3([1:k- 1,k+ 1:n3],:);
        N3= length(C(:,1));
        M3= mean(C,1);s33= cov(C);
        S3= ((n1- 1) * s1+ (n2- 1) * s2+ (N3- 1) * s33)/(n1+ n2+ N3- k);  % 计算混合样本方差
          P01= n1/(n- 1);P02= n2/(n- 1);P03= N3/(n- 1);                   % 计算先验概率
        for i= 1:n3
                W3(i,1)= m1 *inv(S3) *G3(i,:)'- 1/2 *m1 *inv(S3) *m1'+ log(P01);
                W3(i,2)= m2 *inv(S3) *G3(i,:)'- 1/2 *m2 *inv(S3) *m2'+ log(P02);
                W3(i,3)= M3 *inv(S3) *G3(i,:)'- 1/2 *M3 *inv(S3) *M3'+ log(P03); % 计算判别函数
                for j= 1:3
                        if  W3(i,j)= = max(W3(i,:))&j~ = 3
                            N33= N33+ 1;
                        end
                end
        end
end
p11= (N11+ N22+ N33)/(n1+ n2+ n3)
```

输出结果：

```
p11 =
    0.0370
```

由此可见，此题用贝叶斯判别分析效果明显。

习　题　4

1. 已知 $X=(x_1, x_2)^T$ 服从二维正态分布 $N(\mu, \Sigma)$，其中 $\mu=\begin{pmatrix}0\\0\end{pmatrix}$，$\Sigma=\begin{pmatrix}1 & 0.9\\0.9 & 1\end{pmatrix}$，试分别求点 $A=(1, 1)^T$ 和 $B=(1, -1)^T$ 到总体均值的马氏距离和欧氏距离，并论述马氏距离的合理性。

2. 设 G_1、G_2 为两个二维总体，从中分别抽取容量为 3 的训练样本，见表 4-5。

表 4-5　两总体的训练样本

	x_1	x_2		x_1	x_2
G_1	3	2	G_2	6	9
	2	4		5	7
	4	7		4	8

求：1）计算两总体的样本均值向量 $\bar{x}^{(1)}$、$\bar{x}^{(2)}$ 和样本协方差矩阵 S_1、S_2。

2）假定两总体协方差矩阵相等，记为 Σ，用 S_1、S_2 联合估计 Σ。

3）建立距离判别法的判别准则。

4）设有一样品 $x_0=(x_1, x_2)^{\mathrm{T}}=(2, 7)^{\mathrm{T}}$，利用 3）中的判别准则判断其归属。

3. 茶是世界上最为广泛的一种饮料。特选 28 个茶叶样本研究，分别来自江西、云南、福建、广东。按随机次序测试其矿物质元素 Fe、Co、Zn 的含量（单位：$\mu g \cdot g^{-1}$），测试结果见表 4-6，判断四个地区（即四个总体）的协方差矩阵是否相等（$\alpha=0.05$）？

表 4-6　茶叶部分矿物质的测定结果

序　号	产　地	Fe	Co	Zn	序　号	产　地	Fe	Co	Zn
1	江西	244	0.49	48.21	15	福建	153	0.20	11.42
2	江西	141	0.56	44.47	16	福建	175	0.29	14.82
3	江西	147	0.36	43.54	17	福建	245	0.24	10.29
4	江西	147	0.18	54.28	18	福建	171	0.14	10.99
5	云南	478	0.28	39.42	19	福建	256	0.25	11.56
6	云南	497	0.36	54.44	20	福建	160	0.77	35.63
7	云南	494	0.31	47.23	21	福建	195	0.36	30.51
8	云南	345	0.23	32.75	22	福建	216	0.62	29.62
9	福建	188	0.14	27.22	23	福建	146	0.60	21.34
10	福建	171	0.29	11.59	24	福建	228	0.18	38.03
11	福建	199	0.12	12.94	25	福建	476	1.02	35.82
12	福建	159	0.23	11.79	26	广东	169	0.25	16.82

4. 已知两总体的概率密度分别为 $f_1(x)$ 和 $f_2(x)$，且总体的先验分布为 $p_1=0.2$，$p_2=0.8$，误判损失相等。

1）建立贝叶斯判别准则。

2）设有一个新样品 x_0 满足 $f_1(x_0)=6.3$，$f_2(x_0)=0.5$，判定 x_0 的归属。

5. 已知 8 个乳房肿瘤病灶组织的样本见表 4-7，其中前 3 个为良性肿瘤，后 5 个为恶性肿瘤。数据为细胞核显微图像的 5 个量化特征：细胞核直径、质地、周长、面积、光滑度。根据已知样本对未知的 3 个样本进行距离判别和贝叶斯判别，并计算回代误判率与交叉误判率（假定误判损失相等）。

表 4-7　乳房肿瘤病灶组织的样本

序　号	细胞核直径	质　地	周　长	面　积	光 滑 度	类　型
1	13.54	14.36	87.46	566.3	0.097 79	良性
2	13.08	15.71	85.63	520	0.107 5	良性
3	9.504	12.44	60.34	273.9	0.102 4	良性
4	17.99	10.38	122.8	1001	0.118 4	恶性

（续）

序　号	细胞核直径	质　地	周　长	面　积	光 滑 度	类　型
5	20.57	17.77	132.9	1326	0.084 74	恶性
6	19.69	21.25	130	1203	0.109 6	恶性
7	11.42	20.38	77.58	386.1	0.142 5	恶性
8	20.29	14.34	135.1	1297	0.100 3	恶性
9	16.6	28.08	108.3	858.1	0.084 55	待定
10	20.6	29.33	140.1	1265	0.117 8	待定
11	7.76	24.54	47.92	181	0.052 63	待定

实验3　距离判别与贝叶斯判别分析

实验目的

1. 熟练掌握 MATLAB 软件进行距离判别与贝叶斯判别的方法与步骤。
2. 掌握判别分析的回代误判率与交叉误判率的编程。
3. 掌握贝叶斯判别的误判率的计算。

实验数据与内容

我国山区某大型化工厂，在厂区及邻近地区挑选有代表性的 15 个大气取样点，每日 4 次同时抽取大气样品，测定其中含有的 6 种气体的浓度，前后共 4 天，每个取样点每种气体实测 16 次，计算每个取样点每种气体的平均浓度，数据见表 4-8。气体数据对应的污染地区分类见表 4-8 中最后一列。现有两个取自该地区的 4 个气体样本，气体指标见表 4-8 中后 4 行，试解决以下问题：

1. 判别两类总体的协方差矩阵是否相等，然后用马氏距离判别这 4 个未知气体样本的污染类别，并计算回代误判率与交叉误判率；若两类总体服从正态分布，第一类与第二类的先验概率分别为 7/15、8/15，利用贝叶斯判别样本的污染分类。

2. 先验概率为多少时，距离判别与贝叶斯判别相同？调整先验概率对判别结果的影响是什么？

3. 对第一类与第二类的先验概率分别为 7/15、8/15，计算误判概率。

表 4-8　大气样品数据表

气　体	氯	硫 化 氢	二 氧 化 硫	碳　4	环氧氯丙烷	环 己 烷	污染分类
1	0.056	0.084	0.031	0.038	0.0081	0.022	1
2	0.040	0.055	0.100	0.110	0.0220	0.007 3	1
3	0.050	0.074	0.041	0.048	0.007 1	0.020	1
4	0.045	0.050	0.110	0.100	0.025 0	0.006 3	1
5	0.038	0.130	0.079	0.170	0.058 0	0.043	2
6	0.030	0.110	0.070	0.160	0.050 0	0.046	2
7	0.034	0.095	0.058	0.160	0.200	0.029	1

（续）

气　体	氯	硫 化 氢	二氧化硫	碳　4	环氧氯丙烷	环 己 烷	污染分类
8	0.030	0.090	0.068	0.180	0.220	0.039	1
9	0.084	0.066	0.029	0.320	0.012	0.041	2
10	0.085	0.076	0.019	0.300	0.010	0.040	2
11	0.064	0.072	0.020	0.250	0.028	0.038	2
12	0.054	0.065	0.022	0.280	0.021	0.040	2
13	0.048	0.089	0.062	0.260	0.038	0.036	2
14	0.045	0.092	0.072	0.200	0.035	0.032	2
15	0.069	0.087	0.027	0.050	0.089	0.021	1
样品 1	0.052	0.084	0.021	0.037	0.007 1	0.022	待定
样品 2	0.041	0.055	0.110	0.110	0.021 0	0.007 3	待定
样品 3	0.030	0.112	0.072	0.160	0.056	0.021	待定
样品 4	0.074	0.083	0.105	0.190	0.020	1.000	待定

第5章

主成分分析与典型相关分析

在多数实际问题中，往往涉及的数据是多元的统计数据，产生了各种多元统计分析方法。本章介绍主成分分析与典型相关分析这两种多元统计分析方法。主成分分析是利用降维的思想，把多指标转化为少数几个综合指标的一种多元统计分析方法。典型相关分析是研究两组变量间的相关关系，它能够揭示两组变量之间的内在联系，真正反映两组变量间的线性相关情况。

5.1 主成分分析

5.1.1 主成分分析的基本原理

1. 基本思想

在研究实际问题时，往往需要收集多个变量。但这样会使多个变量间存在较强的相关关系，即这些变量间存在较多的信息重复，直接利用它们进行分析，不但模型复杂，还会因为变量间存在多重共线性而引起较大的误差。为能够充分利用数据，通常希望用较少的新变量代替原来较多的旧变量，同时要求这些新变量尽可能反映原变量的信息，这样问题也就简单化了。

主成分分析是采取一种数学降维的方法，找出几个综合变量来代替原来众多的变量，使这些综合变量能尽可能地代表原来变量的信息量，而且彼此之间互不相关。这种把多个变量化为少数几个互相无关的综合变量的统计分析方法就叫做主成分分析或主分量分析。

主成分分析所要做的就是设法将原来众多具有一定相关性的变量，重新组合为一组新的相互无关的综合变量来代替原来变量。通常，数学上的处理方法就是将原来的变量做线性组合，作为新的综合变量，但是这种组合如果不加以限制，则可以有很多，应该如何选择呢？如果将选取的第一个线性组合即第一个综合变量记为 Y_1，自然希望它尽可能多地反映原来变量的信息，这里“信息”用方差来测量，即希望 $\mathrm{var}(Y_1)$越大，表示 Y_1 包含的信息越多。因此在所有的线性组合中所选取的 Y_1 应该是方差最大的，故称 Y_1 为第一主成分。如果第一主成分不足以代表原来 p 个变量的信息，再考虑选取 Y_2 即第二个线性组合，为了有效地反映原来信息，Y_1 已有的信息就不需要再出现在 Y_2 中，用数学语言表达就是要求 $\mathrm{cov}(Y_1, Y_2) = 0$，称 Y_2 为第二主成分，依此类推可以构造出第三、第四……第 p 个主成分。下面介绍这一经典做法的数学原理。

2. 主成分的数学模型

设 $X_1, X_2, \cdots, X_p$ 为实际问题所涉及的 p 个随机变量（可称为 p 项指标），记 $X=(X_1, X_2, \cdots, X_p)^{\mathrm{T}}$，其协方差矩阵为

$$\Sigma=(\sigma_{ij})_p=\mathrm{E}[(\mathrm{X}-\mathrm{E}(\mathrm{X}))(\mathrm{X}-\mathrm{E}(\mathrm{X}))^{\mathrm{T}}]$$

它是一个 p 阶的非负定矩阵。设变量 $x_1,x_2,\cdots,x_p$ 经过线性变换后得到新的综合变量 Y_1，$Y_2,\cdots,Y_p$，即

$$\begin{cases} Y_1 = l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ Y_2 = l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ \cdots \\ Y_p = l_{p1}x_1 + l_{p2}x_2 + \cdots + l_{pp}x_p \end{cases}$$

或

$$Y_i = l_{i1}X_1 + l_{i2}X_2 + \cdots + l_{ip}X_p \quad (i=1,2,\cdots,p) \tag{5.1.1}$$

其中系数 $l_i=(l_{i1},l_{i2},\cdots,l_{ip})$（$i=1,2,\cdots,p$）为常数向量。要求（5.1.1）满足以下条件：

1）系数向量是单位向量，即

$$l_i l_i^{\mathrm{T}} = l_{i1}^2 + l_{i2}^2 + \cdots + l_{ip}^2 = 1 \quad (i=1,2,\cdots,p) \tag{5.1.2}$$

2）Y_i 与 $Y_j(i\neq j,i,j=1,2,\cdots,p)$ 互不相关，即

$$\mathrm{cov}(Y_i,Y_j) = l_i^{\mathrm{T}}\Sigma l_j = 0 \quad (i\neq j,i,j=1,2,\cdots,p) \tag{5.1.3}$$

3）$Y_1,Y_2,\cdots,Y_p$ 的方差递减，即

$$\mathrm{var}(Y_1) \geqslant \mathrm{var}(Y_2) \geqslant \cdots \geqslant \mathrm{var}(Y_p) \geqslant 0 \tag{5.1.4}$$

于是，称 Y_1 为第一主成分，Y_2 为第二主成分，依此类推，有第 p 个主成分。主成分又叫主分量。这里 l_{ij} 称为主成分系数。

3. 主成分的求法及性质

当总体 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的协方差矩阵 $\Sigma=(\sigma_{ij})_p$ 已知时，我们可根据下面的定理求出主成分。

定理 5.1 设协方差矩阵 Σ 的特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant 0$，对应的单位正交特征向量为 $e_1,e_2,\cdots,e_p$，则 X 的第 k 个主成分为

$$Y_k = e_{k1}X_1 + e_{k2}X_2 + \cdots + e_{kp}X_p \quad (k=1,2,\cdots,p) \tag{5.1.5}$$

其中 $e_k=(e_{k1},e_{k2},\cdots,e_{kp})^{\mathrm{T}}$，且

$$\begin{cases} \mathrm{var}(Y_k) = e_k^{\mathrm{T}}\Sigma e_k = \lambda_k \quad (k=1,2,\cdots,p) \\ \mathrm{cov}(Y_k,Y_j) = e_k^{\mathrm{T}}\Sigma e_j = 0 \quad (k\neq j,k,j=1,2,\cdots,p) \end{cases} \tag{5.1.6}$$

证明： 令 $P=(e_1,e_2,\cdots,e_p)$，则为正交矩阵，且

$$P^{\mathrm{T}}\Sigma P = \Lambda = \mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$$

若 $Y_1=l_1^{\mathrm{T}}X$ 为 X 的第一主成分，其中 $l_1^{\mathrm{T}}l_1=1$，令

$$h_1 = (h_{11},h_{12},\cdots,h_{1p}) = P^{\mathrm{T}}l_1$$

则 $h_1^{\mathrm{T}}h_1=1$，$l_1=Ph_1$，且

$$\mathrm{var}(Y_1) = l_1^{\mathrm{T}}\Sigma l_1 = h_1^{\mathrm{T}}P^{\mathrm{T}}\Sigma P h_1 = h_1^{\mathrm{T}}\Lambda h_1 = \lambda_1 h_{11}^2 + \lambda_2 h_{12}^2 + \cdots + \lambda_p h_{1p}^2 \leqslant \lambda_1 h_1^{\mathrm{T}}h_1 = \lambda_1$$

只有当 $h_1=(1,0,\cdots,0)=\varepsilon_1$（标准单位向量）时等号才成立，这时

$$l_1 = Ph_1 = e_1$$

因此，X 的第一主成分为

$$Y_1 = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p$$

且方差 $\mathrm{var}(Y_1)=\lambda_1$ 达到最大。

若 $Y_2=l_2^{\mathrm{T}}X$ 为 X 的第二主成分，其中 $l_2^{\mathrm{T}}l_2=1$，且

$$\mathrm{cov}(Y_1,Y_2)=l_2^{\mathrm{T}}\Sigma e_1=\lambda_1 l_2^{\mathrm{T}}e_1=0$$

令

$$h_2=(h_{21},h_{22},\cdots,h_{2p})=P^{\mathrm{T}}l_2$$

则 $h_2^{\mathrm{T}}h_2=1$，$l_2=Ph_2$，且

$$l_2^{\mathrm{T}}e_1=h_2^{\mathrm{T}}P^{\mathrm{T}}e_1=h_{21}e_1^{\mathrm{T}}e_1+h_{22}e_2^{\mathrm{T}}e_1+\cdots+h_{2p}e_p^{\mathrm{T}}e_1=h_{21}=0$$

从而

$$\mathrm{var}(Y_2)=l_2^{\mathrm{T}}\Sigma l_2=h_2^{\mathrm{T}}P^{\mathrm{T}}\Sigma Ph_2=h_2^{\mathrm{T}}\Lambda h_2=\lambda_1h_{21}^2+\lambda_2h_{22}^2+\cdots+\lambda_ph_{2p}^2=\lambda_2h_{22}^2+\cdots+\lambda_ph_{2p}^2\leqslant\lambda_2h_2^{\mathrm{T}}h_2=\lambda_2$$

只有当 $h_2=(0,1,\cdots,0)=\varepsilon_2$ 时等号才成立，这时

$$l_2=Ph_2=e_2$$

因此，X 的第二主成分为

$$Y_2=e_{21}X_1+e_{22}X_2+\cdots+e_{2p}X_p$$

且方差 $\mathrm{var}(Y_2)=\lambda_2$ 达到最大。

类似可得其余主成分的表达式，且各主成分的方差等于相应的特征值。

定理 5.1 表明：求 X 的主成分等价于求它的协方差矩阵 Σ 的所有特征值及相应的正交单位化特征向量。按特征值由大到小所对应的正交单位化特征向量为组合系数的 $X_1,X_2,\cdots,X_p$ 的线性组合分别为 X 的第一、第二，直至第 P 个主成分，而各主成分的方差等于相应的特征值。

推论　*若记主成分向量* $Y=(Y_1,Y_2,\cdots,Y_p)^{\mathrm{T}}$，*矩阵* $P=(e_1,e_2,\cdots,e_p)$，*则* $Y=P^{\mathrm{T}}X$，*且* Y *的协方差*

$$\Sigma_Y=P^{\mathrm{T}}\Sigma P=\Lambda=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$$

主成分的总方差

$$\sum_{i=1}^{p}\mathrm{var}(Y_i)=\sum_{i=1}^{p}\mathrm{var}(X_i)$$

证明：由（5.1.5）式，显然有 $Y=P^{\mathrm{T}}X$，又由（5.1.6）式，有

$$\sum_{k=1}^{p}\mathrm{var}(Y_k)=\sum_{k=1}^{p}\lambda_k$$

又因为

$$\sum_{k=1}^{p}\lambda_k=\mathrm{tr}(\Sigma_Y)=\mathrm{tr}(\Sigma)=\sum_{k=1}^{p}\mathrm{var}(X_k)$$

所以

$$\sum_{i=1}^{p}\mathrm{var}(Y_i)=\sum_{i=1}^{p}\mathrm{var}(X_i)$$

推论表明：主成分分析可将 p 个原始变量的总方差分解为 p 个不相关变量 $Y_1,Y_2,\cdots,Y_p$ 的方差之和。由于 $\mathrm{var}(Y_k)=\lambda_k(k=1,2,\cdots,p)$，因此，$\lambda_k/\sum_{k=1}^{p}\lambda_k$ 描述了第 k 个主成分提取的信息占总信息的份额。我们称

$$\frac{\mathrm{var}(Y_k)}{\sum_{k=1}^{p}\mathrm{var}(Y_k)}=\frac{\lambda_k}{\sum_{k=1}^{p}\lambda_k}\quad(k=1,2,\cdots,p)\tag{5.1.7}$$

为第 k 个主成分的贡献率，称前 $m(m\leqslant p)$ 个主成分的贡献率之和 $\sum_{k=1}^{m}\lambda_k/\sum_{k=1}^{p}\lambda_k$ 为累计贡献率，它表示前 $m(m\leqslant p)$ 个主成分综合提供总信息的程度。通常选取 $m(m<p)$ 使累计贡献率达到

80%以上。

在实际应用中，选择了重要的主成分后，还要注意主成分实际含义的解释。主成分分析中一个很关键的问题是如何给主成分赋予新的意义，给出合理的解释。一般而言，这个解释是根据主成分表达式的系数结合定性分析来进行的。主成分是原来变量的线性组合，在这个线性组合中变量的系数有大有小，有正有负，有的大小相当，因而不能简单地认为这个主成分是某个原变量的属性的作用。线性组合中各变量系数的绝对值大者表明该主成分主要综合了绝对值大的变量；有几个变量系数大小相当时，应认为这一主成分是这几个变量的总和，这几个变量综合在一起应赋予怎样的实际意义，这要结合具体实际问题和专业，给出恰当的解释，进而才能达到深刻分析的目的。

在 MATLAB 中，运用协方差矩阵进行主成分分析的命令为 pcacov，其调用格式为：

```
1)PC= pcacov(X)
2)[PC,latent,explained]= pcacov(X)
```

其中输入 X 是定理 5.1 中的协方差矩阵；输出 PC 为定理 5.1 中的矩阵 $P=(e_1,e_2,\cdots,e_p)$；latent 是协方差矩阵 Σ 的从大到小排列的特征值向量；explained 表示贡献率向量即每个主成分的方差在观测量总方差中所占的百分数向量。

例 5.1.1 设随机向量 $X=(X_1, X_2, X_3)^T$ 的协方差矩阵为

$$\Sigma=\begin{pmatrix}2 & 2 & -2\\ 2 & 5 & -4\\ -2 & -4 & 5\end{pmatrix}$$

求 X 的各主成分以及各主成分的贡献率。

解：因为已知总体的协方差矩阵，所以可调用主成分分析的命令 pcacov，程序如下：

```
clear
S= [2,2,- 2;2,5,- 4;- 2,- 4,5];          %S 表示总体的协方差矩阵 Σ
[PC,vary,explained]= pcacov(S)           %总体主成分分析
```

程序输出结果：

```
PC =
    - 0.3333      0            0.9428
    - 0.6667      0.7071     - 0.2357
      0.6667      0.7071       0.2357        %主成分变换矩阵
vary=
      10.0000
       1.0000
       1.0000                              %主成分方差向量
explained=
           83.3333
           8.3333
           8.3333                          %各主成分贡献率向量
```

由程序的输出结果以及公式（5.1.5）可知，X 的主成分为：

$$Y_1 = -0.333X_1 - 0.667X_2 + 0.667X_3$$
$$Y_2 = 0X_1 + 0.707X_2 + 0.707X_3$$
$$Y_3 = 0.942X_1 - 0.235X_2 + 0.235X_3$$

主成分的方差

$$\mathrm{var}(Y_1)=\lambda_1=10,\mathrm{var}(Y_2)=\lambda_2=1,\mathrm{var}(Y_3)=\lambda_3=1$$

第一主成分的贡献率

$$\frac{\lambda_1}{\sum_{k=1}^{3}\lambda_k}=83.333\,3\%$$

前两个主成分的累计贡献率为

$$83.333\,3\%+8.333\,3\%=91.666\,6\%$$

因此，若用前两个主成分代替原来三个变量，其信息损失为 8.333 3%，是很小的。

定理 5.2　设 $Y=(Y_1,Y_2,\cdots,Y_p)^{\mathrm{T}}$ 为总体 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的主成分向量，则主成分 Y_i 与变量 X_j 的相关系数

$$\rho_{Y_iX_j}=\frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}e_{ij}\quad(i,j=1,2,\cdots,p)\tag{5.1.8}$$

证明：由定理 5.1 及其推论，因为 $Y=P^{\mathrm{T}}X$，所以 $X=PY$，从而

$$X_j=e_{1j}Y_1+e_{2j}Y_2+\cdots+e_{pj}Y_p$$

于是

$$\mathrm{cov}(Y_i,X_j)=\mathrm{cov}\Big(Y_i,\sum_{i=1}^{p}e_{ij}Y_i\Big)=\lambda_i e_{ij}$$

所以 Y_i 与 X_j 的相关系数为

$$\rho_{Y_iX_j}=\frac{\mathrm{cov}(Y_i,X_j)}{\sqrt{\mathrm{var}(Y_i)}\sqrt{\mathrm{var}(X_j)}}=\frac{\lambda_i e_{ij}}{\sqrt{\lambda_i}\sqrt{\sigma_{jj}}}=\frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}}e_{ij}$$

显然，Y_i 与 X_j 的相关系数反映了主成分 Y_i 与原变量 X_j 的关联程度，它与 X_j 标准差成反比，与主成分的标准差成正比。

若记 $\Sigma_{YX}=(\rho_{Y_iX_j})_p$，则由代数学可以证明：

$$\Sigma_{YX}=(\rho_{Y_iX_j})_p=[\mathrm{diag}(\Sigma)]^{-1/2}P\Lambda^{1/2}\tag{5.1.9}$$

其中 $[\mathrm{diag}(\Sigma)]$ 表示协方差矩阵的主对角线元素组成的对角矩阵，P 是主成分矩阵，Λ 是特征值对角矩阵。

例 5.1.2（续例 5.1.1）　求主成分与原变量的相关系数。

解：由例 5.1.1 的条件与结果，根据（5.1.9）式，可编写 MATLAB 程序如下：

```
clear
S=[2,2,-2;2,5,-4;-2,-4,5];                %S 表示总体的协方差矩阵 Σ
[PC,vary,explained]=pcacov(S);            %总体主成分分析
S1=diag(diag(S));                         %协方差矩阵的主对角线元组成的对角矩阵
SYX=inv(sqrt(S1))*PC*sqrt(diag(vary));    %按(5.1.9)式计算
SYX
```

程序输出结果：

```
SYX =
    -0.7454    0          0.6667
    -0.9428    0.3162    -0.1054
     0.9428    0.3162     0.1054
```

所以，SYX 的第一列元依次为 Y_1 与 X_1、X_2、X_3 的相关系数，即 $\rho_{Y_1X_1}=-0.745\,4$，

$\rho_{Y_1X_2}=-0.9428$，$\rho_{Y_1X_3}=0.9428$，其余各列的元类推。结果表明 Y_1 与 X_2、X_3 高度相关，Y_2 与 X_1 不相关等。

4. 标准化变量的主成分

在解决实际问题的过程中，经常遇到不同的指标具有不同的量纲，有时会导致各指标取值的分散程度较大，这样在计算协方差矩阵时，可能出现总体的方差主要受方差较大的数据的控制，可能造成不合理的结果。为了消除量纲的影响，通常对原始数据进行标准化，令

$$X_i^* = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}} \quad (i=1,2,\cdots,p) \tag{5.1.10}$$

其中 $\mu_i=EX_i$，$\sigma_{ii}=\mathrm{var}(X_i)$，这时 $X^*=(X_1^*,X_2^*,\cdots,X_p^*)^{\mathrm{T}}$ 的协方差矩阵就是原始数据的相关系数矩阵 $\rho=(\rho_{ij})_{p\times p}$，其中

$$\rho_{ij}=\frac{\mathrm{cov}(X_i,X_j)}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

计算标准化变量的主成分公式为：

$$Y_i^* = (e_i^*)^{\mathrm{T}}X^* = e_{i1}^*X_1^* + e_{i2}^*X_2^* + \cdots + e_{ip}^*X_p^* \quad (i=1,2,\cdots,p) \tag{5.1.11}$$

其中 X^* 是标准化以后的数据，$(e_i^*)^{\mathrm{T}}$ 是相关系数矩阵的特征值对应的特征向量。

标准化变量的主成分具有以下性质。

性质 1 总体方差和等于向量的维数，即

$$\sum_{i=1}^{p}\mathrm{var}(Y_i^*)=\sum_{i=1}^{p}\mathrm{var}(X_i^*)=\sum_{i=1}^{p}\lambda_i^*=p$$

其中 $\lambda_i^*(i=1,2,\cdots,p)$ 是相关系数矩阵 ρ 的特征值。

性质 2 标准化变量的第 i 个主成分的贡献率为

$$\lambda_i^*/p \quad (i=1,2,\cdots,p)$$

标准化变量的前 m 个主成分的累积贡献率为

$$\sum_{i=1}^{m}\lambda_i^*/p$$

性质 3 主成分 Y_i^* 与标准化数据 X_j^* 的相关系数为

$$\rho(Y_i^*,X_j^*)=\sqrt{\lambda_i^*}e_{ij}^*$$

值得注意的是：同一个总体，分别从协方差矩阵和相关系数矩阵出发进行主成分分析，所得的主成分的贡献率可以不同，读者可参考本章习题的第 1 题。

5.1.2 样本主成分分析

实际问题中，总体 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的协方差矩阵 Σ 一般是未知的，具有的资料只是来自于 X 的一个容量为 n 的样本观测数据。设

$$x_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}} \quad (i=1,2,\cdots,n)$$

为取自总体 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$ 的一个容量为 n 的简单随机样本，由第 2 章知，样本协方差矩阵及样本相关矩阵分别为

$$S=(s_{ij})_{p\times p}=\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})^{\mathrm{T}}$$

$$R=(r_{ij})_{p\times p}=\left(\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}\right)$$

其中 $\bar{x}=(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_p)^{\mathrm{T}}$，$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$，$s_{jk}=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)(j,k=1,2,\cdots,p)$。

分别以 S 和 R 作为总体 Σ 和 ρ 的估计，然后按总体主成分分析的方法作样本主成分分析。

关于样本主成分，有如下结论。

设 $S_{p\times p}$ 为样本协方差矩阵，其特征值 $\hat{\lambda}_1\geqslant\hat{\lambda}_2\geqslant\cdots\geqslant\hat{\lambda}_p\geqslant 0$，相应的单位正交化特征向量 $\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_p$，第 k 个样本主成分为

$$y_k=\hat{e}_k^{\mathrm{T}}x=\hat{e}_{k1}x_1+\hat{e}_{k2}x_2+\cdots+\hat{e}_{kp}x_p$$

当依次代入观测值 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}(i=1,2,\cdots,n)$ 时，便得到第 k 个样本主成分 y_k 的 n 个观测值

$$y_{1k},y_{2k},\cdots,y_{nk}$$

称其为第 k 个样本主成分的得分。这时

$$\begin{cases}y_k\text{ 得分的样本方差}=\hat{e}_k^{\mathrm{T}}S\,\hat{e}_k=\hat{\lambda}_k(k=1,2,\cdots,p)\\ y_j\text{ 与 }y_k\text{ 得分的协方差}=\hat{e}_j^{\mathrm{T}}S\,\hat{e}_k=0(j\neq k)\\ \text{样本总方差}=\sum_{k=1}^{p}s_{kk}=\sum_{k=1}^{p}\hat{\lambda}_k\end{cases}$$

第 k 个主成分 Y_k 的贡献率 $\hat{\lambda}_k/\sum_{i=1}^{p}\hat{\lambda}_i$，前 m 个样本主成分的累计贡献率为 $\sum_{k=1}^{m}\hat{\lambda}_k/\sum_{i=1}^{p}\hat{\lambda}_i$。

同总体主成分分析一样，为了消除量纲的影响，可利用第 2 章介绍的方法对样本进行标准化，即令

$$x_i^{*}=\left(\frac{x_{i1}-\bar{x}_1}{\sqrt{s_{11}}},\frac{x_{i2}-\bar{x}_2}{\sqrt{s_{22}}},\cdots,\frac{x_{ip}-\bar{x}_p}{\sqrt{s_{pp}}}\right)^{\mathrm{T}}\quad(i=1,2,\cdots,n)\tag{5.1.12}$$

则标准化的样本数据的协方差矩阵即原始数据的样本相关系数矩阵，由样本相关系数矩阵出发作主成分分析即可。

在 MATLAB 中，运用样本数据矩阵进行主成分分析的命令为 princomp，其调用格式为：

```
1)PC= princomp(X)
2)[PC,SCORE,latent,tsquare]= princomp(X)
```

其中输入 X 是样本数据矩阵；输出 PC 为主成分矩阵 $P=(\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_p)$；SCORE 是样本主成分的得分；latent 是 X 的方差矩阵的特征值向量；tsquare 是每个数据点的 HotellingT2 统计量。

综上，由样本观测数据矩阵进行主成分分析的步骤为：

第一步：对原始数据进行标准化处理；

第二步：计算样本相关系数矩阵；

第三步：求相关系数矩阵的特征值和相应的特征向量；

第四步：选择重要的主成分，并写出主成分表达式；

第五步：计算主成分得分；

第六步：依据主成分得分的数据，进行进一步的统计分析。

例 5.1.3　对 10 名男中学生的身高（X_1）、胸围（X_2）和体重（X_3）进行测量，得数据见表 5-1，对其做主成分分析。

表 5-1　10 名男中学生的身高、胸围及体重数据

序　号	身高 X_1(cm)	胸围 X_2(cm)	体重 X_3(kg)	序　号	身高 X_1(cm)	胸围 X_2(cm)	体重 X_3(kg)
1	149.5	69.5	38.5	6	156.1	74.5	45.5
2	162.5	77	55.5	7	172.0	76.5	51.0
3	162.7	78.5	50.8	8	173.2	81.5	59.5
4	162.2	87.5	65.5	9	159.5	74.5	43.5
5	156.5	74.5	49.0	10	157.7	79	53.5

解：令 $X=(X_1, X_2, X_3)$ 表示总体，则表 5-1 中的数据可认作为来自总体的样本，由样本主成分分析法，MATLAB 程序如下：

```
X=[149.5  69.5  38.5;162.5  77  55.5;162.7  78.5  50.8;162.2  87.5  65.5;156.5  74.5
49.0;156.1  74.5  45.5;172.0  76.5  51.0;173.2  81.5  59.5;159.5  74.5  43.5;157.7  79
53.5];                                    %输入样本数据
[P,SCORE,latent]= princomp(X)             %主成分分析
```

程序输出结果为：

```
P =
    0.5592    0.8277  - 0.0480
    0.4213  - 0.3335  - 0.8434
    0.7140  - 0.4514    0.5352              %正交单位化特征向量
latent=
        110.0041
        25.3245
        1.5680                              %特征值
```

所以，样本各主成分的贡献率分别为

$$\frac{110.004}{136.896}=80.36\%,\quad \frac{25.324}{136.896}=18.50\%,\quad \frac{1.568}{136.896}=1.15\%$$

因此，前两个主成分的累计贡献率已达 98.860%，实际应用中可只取前两个主成分，即

$$y_1 = 0.559\,2x_1 + 0.421\,3x_2 + 0.714\,0x_3$$

$$y_2 = 0.827\,7x_1 - 0.333\,5x_2 - 0.451\,4x_3$$

其中 (x_1, x_2, x_3) 是 (X_1, X_2, X_3)。

第一主成分 y_1 是身高值 x_1、胸围值 x_2 和体重值 x_3 的加权和，当一个学生的值 y_1 较大时，可以推断他较高或较胖或又高又胖；反之，当一个学生比较魁梧时，所对应的 y_1 值一般也较大，故第一主成分是反映学生身材是否魁梧的综合指标，我们一般称之为“大小”因子。而在第二主成分的表达式中，身高 x_1 前的系数为正，而胸围 x_2 和体重 x_3 前的系数为负，当一个学生的 y_2 值较大时，说明 x_1 的值较大，而 x_2 和 x_3 的值相对较小，即该生较高且瘦；反之，瘦高型学生的 y_2 值会较大，故 y_2 是反映学生体型特征的综合指标，我们一般称之为“形状”因子。

例 5.1.4　根据调查分析，影响我国粮食安全生产的主要因素有以下几个方面：有效灌溉面积（X_1）、粮食播种面积（X_2）、成灾面积（X_3）、财政投入（X_4）、农业劳动力（X_5）、农村用电量（X_6）、农业机械总动力（X_7）及农业化肥施用量（X_8），具体数据见表 5-2。对粮食安全生产因素作主成分分析。

表 5-2　影响我国粮食安全生产的主要因素表

年份	有效灌溉面积（万公顷）	粮食播种面积（万公顷）	成灾面积（万公顷）	财政投入（万元）	农业劳动力（万人）	农村用电量（万千瓦）	农机总动力（万千瓦）	化肥施用量（万千千克）
1990	4 740.31	11 346.60	178.20	221.76	33 336.00	844.50	28 707.70	647.58
1991	4 782.21	11 231.40	278.10	243.55	34 186.30	963.20	29 388.60	701.28
1992	4 859.01	11 056.00	259.00	269.04	34 037.00	1 106.90	30 308.40	732.55
1993	4 872.79	11 050.90	231.30	323.42	33 258.20	1 244.80	31 816.60	787.98
1994	4 875.91	10 854.40	313.80	399.70	32 690.30	1 473.90	33 802.50	829.53
1995	4 928.12	11 006.00	222.70	430.22	32 335.00	1 655.70	36 118.10	898.40
1996	5 038.14	11 254.80	212.30	510.07	32 260.40	1 812.70	38 546.90	957.00
1997	5 123.85	11 291.20	303.10	560.77	32 434.90	1 980.10	42 015.60	995.23
1998	5 229.56	11 378.70	251.80	626.02	32 626.40	2 042.10	45 207.70	1 020.88
1999	5 315.80	11 316.10	267.30	677.46	32 911.80	2 173.40	48 996.10	1 031.08
2000	5 382.00	10 846.30	343.70	766.89	32 798.00	2 421.30	52 573.60	1 036.63
2001	5 424.90	10 608.00	317.90	917.96	32 451.00	2 610.80	55 172.10	1 063.28
2002	5 435.50	10 389.10	271.60	1 102.70	31 991.00	2 993.40	57 929.90	1 083.08
2003	5 401.42	9 941.00	325.20	1 134.86	31 259.60	3 432.90	60 386.50	1 102.90
2004	5 447.80	10 160.60	163.00	1 693.79	30 596.00	3 933.00	64 027.90	1 157.30
2005	5 502.93	10 427.80	199.70	1 792.40	29 975.50	4 375.70	68 397.80	1 191.45
2006	5 575.05	10 495.80	246.30	2 161.35	28 886.35	4 895.80	72 522.10	1 231.90
2007	5 651.83	10 563.80	250.60	3 404.70	22 543.4	5 509.90	76 589.60	1 276.70
2008	5 847.17	10 679.30	222.80	4 544.01	20 078.6	5 713.20	82 190.41	1 309.19

资料来源：《中国统计年鉴 2008》。

解：由于各个指标的单位不同，且各指标的方差相差很大，所以首先对样本数据进行无量纲的变换，变换方法是用采用标准化方法。然后对标准化的样本数据进行主成分分析。程序如下：

```
%将表 5-2 中各指标的观测值作为矩阵 X 输入,省略号表示书写时数据省略了
X= [4740.31, 11346.60, 178.20, 221.76, 33336.00, 844.50, 28707.70, 647.58;
   ...
   5847.17, 10679.30, 222.80, 4544.01, 20078.6, 5713.20, 82190.41, 1309.19];
X1= zscore(X);                    %按公式(5.1.12)对样本数据标准化
[pc,la,tent]= princomp(X1);       %主成分分析,pc 是特征向量矩阵,la 是得分矩阵,tent 是特征值
tents= sum(tent);                 %特征值总和
gxl=  tent/ tents;                %各个主成分贡献率
```

程序输出结果：

```
pc =                                          %正交单位化特征向量
    0.3933   - 0.1518   - 0.0544    0.3944    0.5494    0.3701   - 0.1396    0.4533
  - 0.2821    0.3173   - 0.7669    0.4403    0.0325  - 0.1907   - 0.0141  - 0.0121
  - 0.0479  - 0.8701   - 0.4379  - 0.1998  - 0.0337  - 0.0838    0.0257  - 0.0115
    0.3856    0.1952   - 0.2572  - 0.3364    0.2998    0.0210    0.7181  - 0.1668
  - 0.3605  - 0.2487    0.3850    0.4799    0.2640  - 0.3102    0.5109  - 0.0512
    0.4083    0.0358    0.0265  - 0.0014  - 0.2123  - 0.7233    0.0088    0.5128
    0.4070  - 0.0643    0.0422    0.1814    0.2707  - 0.3433  - 0.3636  - 0.6872
    0.3905  - 0.1175  - 0.0151    0.4841  - 0.6466    0.2861    0.2659  - 0.1692
```

```
tent=
    5.9106
    1.1327
    0.6118
    0.2842
    0.0347
    0.0216
    0.0039
    0.0005                          %特征值
gxl =
    0.7388
    0.1416
    0.0765
    0.0355
    0.0043
    0.0027
    0.0005
    0.0001                          %样本各主成分的贡献率
```

所以，第一、二主成分的累计贡献率为

$$0.738\,8+0.141\,6=0.880\,4=88.04\%$$

可取前两个主成分，即

$$y_1=e_1^{\mathrm{T}}(x_1^*,x_2^*,x_3^*,x_4^*,x_5^*,x_6^*,x_7^*,x_8^*)^{\mathrm{T}}$$

$$y_2=e_2^{\mathrm{T}}(x_1^*,x_2^*,x_3^*,x_4^*,x_5^*,x_6^*,x_7^*,x_8^*)^{\mathrm{T}}$$

其中 $e_1^{\mathrm{T}}=(0.393\,3,\ -0.282\,1,\ -0.047\,9,\ 0.385\,6,\ -0.360\,5,\ 0.408\,3,\ 0.407\,0,\ 0.390\,5)$，$e_2^{\mathrm{T}}=(-0.151\,8,\ 0.317\,3,\ -0.870\,1,\ 0.195\,2,\ -0.248\,7,\ 0.035\,8,\ -0.064\,3,\ -0.117\,5)$，$x_i^*$ 由（5.1.12）式定义。

第一主成分 y_1 反映了我国农业生产基础设施投入的情况，第二主成分 y_2 反映了我国粮食生产抗灾能力与劳动力情况。

需要指出的是，关于主成分的实际意义，要结合具体问题和有关的专业知识才能给出合理的解释。虽然利用主成分本身可对所研究的问题在一定程度上做分析，但主成分分析往往并不是最终目的，通常是将主成分综合原始数据的信息，达到降低原始数据维数的目的，进而将少数几个主成分的得分作为新数据，对其再做进一步分析，如基于主成分的回归分析、聚类分析等。

5.2　主成分分析的应用

主成分分析的应用范围非常广泛，诸如投资组合风险管理、企业效益的综合评价、图像特征识别、机械加工或传感器故障检测、灾害损失分析等。将主成分分析与聚类分析、判别分析以及回归分析方法相结合，还可以解决更多实际问题。

5.2.1　主成分分析用于综合评价

主成分分析用于综合评价的一般步骤：

1）若各指标的属性不同（成本型、利润型、适度型），则将原始数据矩阵 A 统一趋势化，得到属性一致的指标矩阵 B（具体过程参见第 2 章的数据变换一节）。

2）计算 B 的协方差矩阵 Σ，或相关系数矩阵 R（当 B 的量纲不同或 Σ 矩阵主对角元素差距过大时，用相关系数矩阵 R）。

3）计算 Σ 或 R 的特征值与相应的特征向量。

4）根据特征值计算累计贡献率，确定主成分的个数，而特征向量 V 就是主成分的系数向量。

5）计算主成分的数值（即主成分得分）。若利用协方差矩阵 Σ 计算特征值与特征向量，则主成分得分为

$$F = (B - EB)^* V$$

若利用相关系数矩阵 R 计算特征值与特征向量，则主成分得分为：

$$F = B^* V$$

其中，V 是特征向量矩阵，B^* 是将矩阵 B 标准化以后的矩阵（即 zscore（B））。

6）计算综合评价值，进行排序。若为效益型矩阵，则评价值越大排名越靠前；若为成本型矩阵，则评价值越小排名越靠前。

通常计算综合评价值的公式为

$$Z = FW$$

其中 F 是主成分得分矩阵，W 是将特征值归一化后得到的权向量。

例 5.2.1　根据 2008 年安徽统计年鉴资料，选择 x_1（工业总产值的现价）、x_2（工业销售按当年价的产值）、x_3（流动资产年平均余额）、x_4（固定资产净值年平均余额）、x_5（业务收入）、x_6（利润总额）6 项指标进行主成分分析，表 5-3 列出了安徽省各市大中型工业企业主要经济指标的统计数据。(1) 选取指标是否合适？(2) 给出各市大中型工业企业排名。

表 5-3　安徽省各市大中型工业企业主要经济指标　　（单位：亿元）

地　区	x_1	x_2	x_3	x_4	x_5	x_6
合肥市	1 932.27	1 900.53	653.83	570.95	1 810.70	119.53
淮北市	367.05	366.08	186.16	252.07	395.43	32.82
亳州市	86.89	85.38	40.85	51.71	83.26	8.95
宿州市	154.27	147.07	30.68	57.96	146.30	−1.27
蚌埠市	197.21	193.28	104.56	90.15	182.60	7.85
阜阳市	244.17	231.55	56.37	121.96	224.04	26.49
淮南市	497.74	483.69	206.80	501.37	496.59	27.76
滁州市	308.91	296.99	118.65	76.90	277.42	19.32
六安市	191.77	189.05	70.19	62.31	191.98	23.08
马鞍山市	905.32	894.61	351.52	502.99	1 048.02	53.88
巢湖市	254.99	242.38	106.66	75.48	234.76	19.65
芜湖市	867.07	852.34	418.82	217.76	806.94	37.01
宣城市	219.36	207.07	82.58	54.74	192.74	11.02
铜陵市	570.33	563.33	224.23	190.77	697.91	20.61
池州市	59.11	57.32	16.97	40.33	56.56	6.03
安庆市	430.58	426.25	103.08	147.05	442.04	0.79
黄山市	65.03	64.36	28.38	8.58	60.48	2.88

数据来源：《安徽统计年鉴 2008》。

解：首先输入数据，程序如下：

```
A= [data];                          %data 即表 5-3 中数据
R= corrcoef(A);
```

得到的相关系数矩阵为：

$$R=\begin{pmatrix}1.000\,0 & 1.000\,0 & 0.975\,4 & 0.823\,1 & 0.991\,4 & 0.937\,5\\ 1.000\,0 & 1.000\,0 & 0.975\,8 & 0.823\,6 & 0.992\,0 & 0.936\,9\\ 0.975\,4 & 0.975\,8 & 1.000\,0 & 0.824\,5 & 0.971\,2 & 0.912\,7\\ 0.823\,1 & 0.823\,6 & 0.824\,5 & 1.000\,0 & 0.850\,2 & 0.802\,0\\ 0.991\,4 & 0.992\,0 & 0.971\,2 & 0.850\,2 & 1.000\,0 & 0.921\,2\\ 0.937\,5 & 0.936\,9 & 0.912\,7 & 0.802\,0 & 0.921\,2 & 1.000\,0\end{pmatrix}$$

由于 $r_{12}=r_{21}=1$，表明指标 x_1、x_2 完全线性相关，故只需保留一个指标，程序如下：

```
A1= A(:,2:6)./[ones(17,1)*std(A(:,2:6))];      %消除量纲
[v,d]= eig(corrcoef(A1));                      %计算特征值与特征向量
w= sum(d)/sum(sum(d));                         %计算贡献率
F= [A1- ones(17,1)*mean(A1)]*d(:,5);           %计算主成分得分
[F1,I1]= sort(F,'descend');                    %I1 给出各名次的序号
[F2,I2]= sort(I1);                             %I2 给出各市排名
```

计算结果见表 5-4 与表 5-5。

表 5-4　特征值、特征向量及贡献率

特　征　值	特 征 向 量	贡　献　率
4.610 0	(0.459 5，0.455 2，0.415 8，0.460 0，0.444 1)	0.922 0
0.247 5	(−0.251 7，−0.210 3，0.905 4，−0.131 5，−0.235 4)	0.049 5
0.105 0	(0.192 6，0.370 2，−0.039 0，0.302 9，−0.855 9)	0.021 0
0.032 2	(−0.351 0，0.777 9，0.027 5，−0.515 3，0.073 8)	0.006 4
0.005 3	(0.751 8，−0.080 3，0.071 9，−0.643 4，−0.096 5)	0.001 1

表 5-5　各市第一主成分得分排名

地区	得分	排名	地区	得分	排名	地区	得分	排名
合肥	15.407 5	1	淮南	0.529 5	5	宣城	−2.184 5	11
淮北	1.349 8	4	滁州	−0.838 8	10	铜陵	−0.629 7	8
亳州	−2.520 1	12	六安	−0.229 3	7	池州	−2.993 5	14
宿州	−4.176 9	17	马鞍山	4.764 1	2	安庆	−3.843 0	16
蚌埠	−2.698 4	13	巢湖	−0.785 3	9	黄山	−3.504 1	15
阜阳	0.323 6	6	芜湖	2.029 1	3			

5.2.2　主成分分析用于分类

利用主成分分析可以计算出主成分的得分，如果主成分的贡献率较大，则其提取了原始数据的主要信息，因此可以利用主成分分析进行分类。

例 5.2.2　做出例 4.1.1 中蠓虫原始数据图与主成分得分数据图。

解：作图程序如下：

```
clear
apf= [1.14,1.78;1.18,1.96;1.20,1.86;1.26,2.00;1.28,2.00;1.30,1.96];
af= [1.24,1.72;1.36,1.74;1.38,1.64;1.38,1.82;1.38,1.90;1.40,1.70;1.48,1.82;1.54,1.82;1.56,
2.08];
x= [1.24,1.8;1.28,1.84; 1.4,2.04];                                    %输入原始数据

subplot(2,1,1)
plot(apf(:,1),apf(:,2),'*',af(:,1),af(:,2),'or',x(:,1),x(:,2),'p')    %原始数据图形

[c1,s1,l1,t1]= princomp([apf;af;x]);                                  %计算主成分得分 s1

subplot(2,1,2),
plot(1:6,s1(1:6,2),'*')

hold on
plot(7:15,s1(7:15,2),'or')

hold on
plot(16:18,s1(16:18,2),'p')
legend('apf','af','x')                                                %主成分得分图形

hold on
plot(0:18,0*ones(1,19),'- ')                                          %画直线
```

从图 5-1 上方图形可以看出原始数据的图形中不同蠓虫触长与翅长两个指标无法做到楚汉分明，但是下方主成分得分的图形则显示 Af 类蠓虫的第一个主成分大于 Apf 的得分，而第二主成分得分小于 Apf 的得分，且基本小于零；未知的 3 只蠓虫第一主成分的得分最大，第二主成分的得分大于 Af 而小于 Apf。依此为判据可以得到与第 4 章马氏距离判别同样的结果。

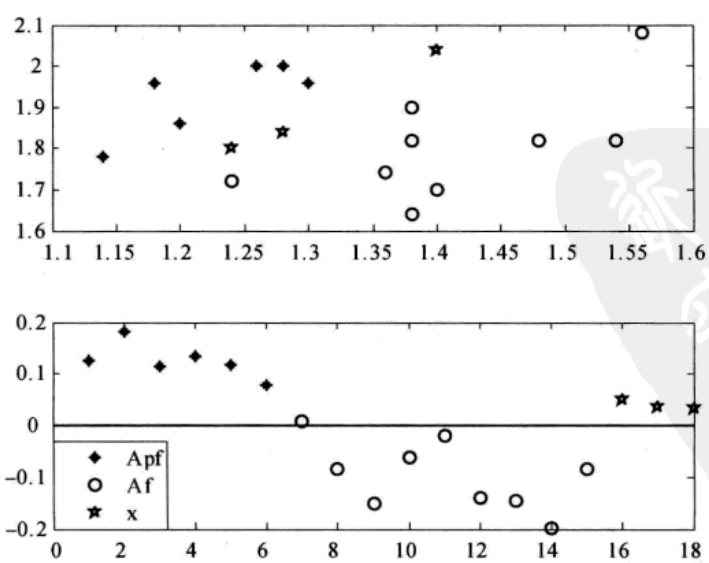


图 5-1　蠓虫原始数据图（上方）与主成分得分数据图（下方）

由于例 4.1.1 蠓虫的原始数据是二维向量，可以做出图形，如果原始数据维数大于 3，借助主成分分析，我们可以选取 2～3 个主成分，做出主成分得分的散点图。

例 5.2.3 瑞士银行纸币（见附录：瑞士银行纸币（Swiss Bank Notes））数据为 200×6 的矩阵，其中前 100 行是真纸币数据，后 100 行是假纸币数据。六项指标为：纸币长度，左、右侧纸币高度，上、下图廓内骨架边距以及对角线长度，分别用 $X_1, X_2, \cdots, X_6$ 表示。（1）利用协方差矩阵进行主成分分析，此时可否利用第一主成分得分进行排名？（2）利用 R 矩阵进行主成分分析，此时可否利用第一主成分得分进行排名？（3）选择两个主成分的得分做出平面图形，能否从图形上分别真假纸币？

解：（1）首先输入原始数据，然后利用协方差矩阵进行主成分分析，程序如下：

```
a= [data];                          %输入原始数据
[M,N]= size(a);
[v,d]= eig(cov(a));                 %样本协方差矩阵的特征值
```

输出结果显示，最大特征值对应的不是正向量，所以不能用第一主成分得分进行排名。

（2）利用 R 矩阵进行主成分分析，程序如下：

```
a= [data];                                                    %输入原始数据
[M,N]= size(a);                                               %计算原始数据维数
for i= 1:N
    for j= 1:N
      R(i,j)= 2*dot(a(:,i),a(:,j))./[sum(a(:,i).^2)+ sum(a(:,j).^2)];    %计算 R 矩阵
    end
end
[v,d]= eig(R);                                                %R 矩阵的特征值与特征
向量
q= sum(d)/sum(sum(d));                                        %计算贡献率
```

输出结果显示，最大特征值对应的是正向量，且其贡献率达到 60%，所以可以用第一主成分得分进行排名。

（3）分别选择第一、第二两个主成分，第二、第三主成分以及第一、第三两个主成分的得分做出平面图形，程序如下：

```
a= [data];                                                    %输入原始数据
[M,N]= size(a);
[v,d]= eig(cov(a));                                           %样本协方差矩阵的特征值
q= sum(d)/sum(sum(d));                                        %计算贡献率
F= a*v;                                                       %计算主成分得分
subplot(2,2,1)
plot(F(1:100,6),F(1:100,5),'o',F(101:200,6),F(101:200,5),'+ ')      %第一、二两个主成分
title('w- pc1- pc2')
subplot(2,2,2)
plot(F(1:100,5),F(1:100,4),'or',F(101:200,5),F(101:200,4),'+ ')     %第二、三两个主成分
title('w- pc2- pc3')
subplot(2,2,3)
plot(F(1:100,6),F(1:100,4),'or',F(101:200,6),F(101:200,4),'+ ')     %第一、三两个主成分
title('w- pc1- pc3')
```

```
subplot(2,2,4)
plot(1:6,fliplr(sum(d)),'- or')                    %从大到小特征值
legend('eigenvalues ')
title('从大到小特征值')
```

输出如图 5-2 所示。从图 5-2 可以看出，第一、第二两个主成分以及第一、第三两个主成分的得分做出平面图形还可以区分真假纸币，而第二、第三两个主成分的得分做出的图形（图 5-2 右上）则呈现出混乱现象。为了使得从图形上更好地区别真假纸币，我们可以计算加权主成分得分，将上面程序中计算主成分得分的命令行“F= a*v”改写成如下形式：

```
w= q/sum(q);                                       %计算权向量
F= [a.*(ones(M,1)*w)]*v;                           %计算加权主成分得分
```

绘图命令如“title（'pc1-pc2'）改写成 title（'w-pc1-pc2'），加了字符“w-”，此时输出图形如图 5-3 所示，其效果明显优于图 5-2。

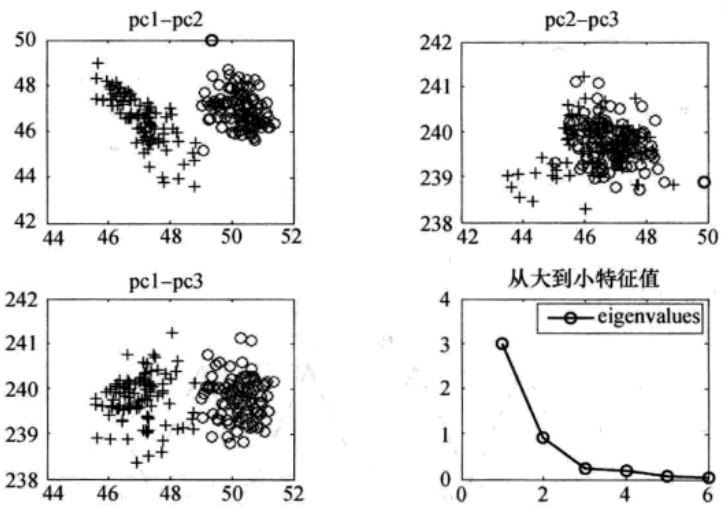


图 5-2　瑞士银行纸币主成分图形

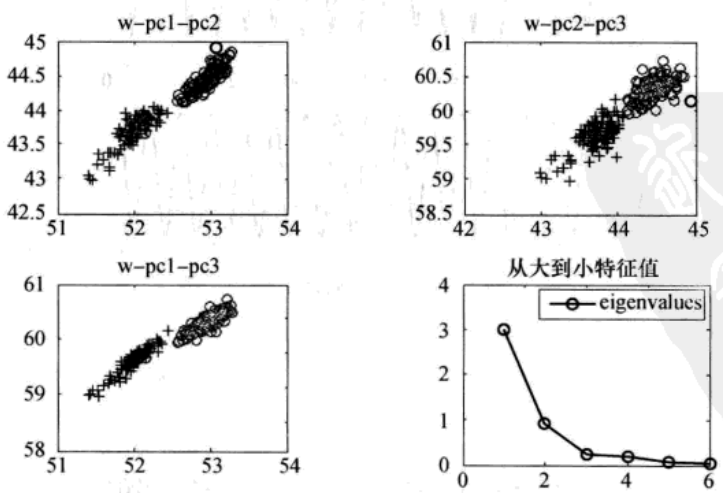


图 5-3　瑞士银行纸币加权主成分图形

5.2.3　主成分分析用于信号分离

在电波、声波信号传输过程中由于各种干扰较多，给实际数据带来了一定的噪声污染，主成分分析利用真实信号在混合信号中占主导地位的特性，可以进行信号分离与提取。

例 5.2.4　设 $x_1=\sin\frac{t}{31}$，$x_2=|\cos(1.89t)|$，$x_3=\sin(3.43t)$，$t\in[1,\ 10\pi]$，(1) 构造信号 $s_1=x_1-\overline{x}_1$，$s_2=x_2-\overline{x}_2$，$s_3=x_3$，并做出图像；(2) 利用 MZ 算法 (Marsaglia's Ziggurat algorithm) 生成随机噪声 z_t，将 3 个信号混合为 $y_1=0.1s_1+0.8s_2+0.01z_t$，$y_2=0.4s_1+0.3s_2+0.01z_t$，$y_3=0.1s_1+s_3+0.02z_t$；(3) 使用主成分分析与独立成分分析将混合信号分离。

解：(1) 构造信号的程序如下：

```
clear
i = [1:0.01:10 * pi]';                                    %输入 t 的取值
[dummy index] = sort(sin(i));
s1(index,1) = i/31; s1 = s1 - mean(s1);                   %生成 s1
s2 = abs(cos(1.89*i)); s2 = s2 - mean(s2);                %生成 s2
s3 = sin(3.43*i);                                         %生成 s3
subplot(311), plot(s1), ylabel('s_1'), title('Raw signals')
subplot(312), plot(s2), ylabel('s_2')
subplot(313), plot(s3), ylabel('s_3')
```

运行结果如图 5-4 所示。

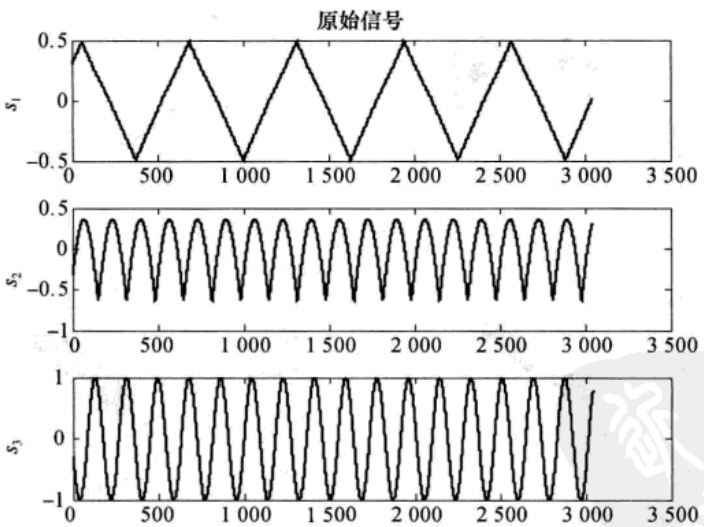


图 5-4　3 个原始信号图形

(2) 生成 3 个混合信号的程序如下：

```
randn('state',1);                                         %生成随机噪声
y1= 0.1*s1 + 0.8*s2 + 0.01*randn(length(i),1);            %生成混合信号 y1
y2= 0.4*s1 + 0.3*s2 + 0.01*randn(length(i),1);            %生成混合信号 y2
y3= 0.1*s1 + s3+ 0.02*randn(length(i),1);                 %生成混合信号 y3
y= [y1,y2,y3];
subplot(311), plot(y(:,1)), ylabel('y_1'), title('Mixed signals')
```

```
subplot(312), plot(y(:,2)), ylabel('y_2')
subplot(313), plot(y(:,3)), ylabel('y_3')
```

运行结果如图 5-5 所示。

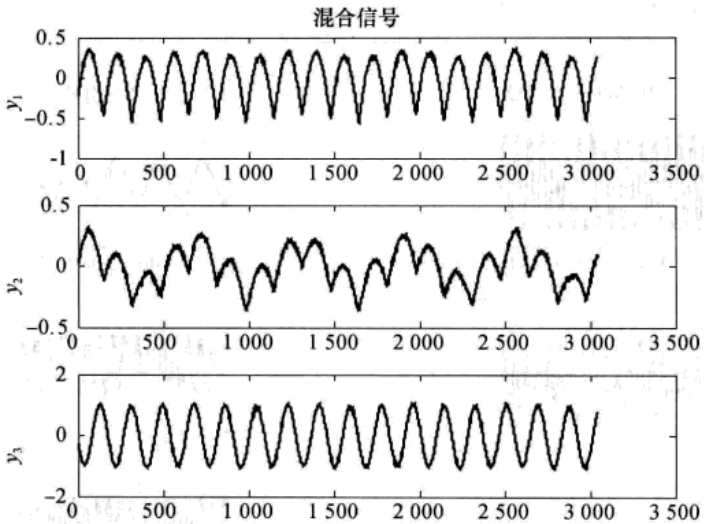


图 5-5　混合信号图形

(3) 分离信号的程序如下：

```
%%主成分分析
[coeff, score,latent] = princomp(y)
sPCA= score;
sPCA= sPCA./repmat(std(sPCA),length(sPCA),1);
subplot(3,2,1)
plot(sPCA(:,1))
ylabel('s_{PCA1}'), title('Separated signals - PCA')
subplot(3,2,3)
plot(sPCA(:,2)), ylabel('s_{PCA2}')
subplot(3,2,5)
plot(sPCA(:,3)), ylabel('s_{PCA3}')
%独立成分分析
rand('state',1);
div= 0;
B= orth(rand(3, 3)- 0.5);
BOld = zeros(size(B));
while (1- div) > eps
   B= B*real(inv(B'*B)^(1/2));
   div= min(abs(diag(B'*BOld)));
   BOld= B;
   B= (sPCA'*(sPCA*B).^3)/length(sPCA)- 3*B;
   sICA= sPCA*B;
end
subplot(3,2,2)
plot(sICA(:,1)), ylabel('s_{ICA1}'), title('Separated signals - ICA')
```

```
subplot(3,2,4)
plot(sICA(:,2)), ylabel('s_{ICA2}')
subplot(3,2,6)
plot(sICA(:,3)), ylabel('s_{ICA3}')
```

运行结果如图 5-6 所示。

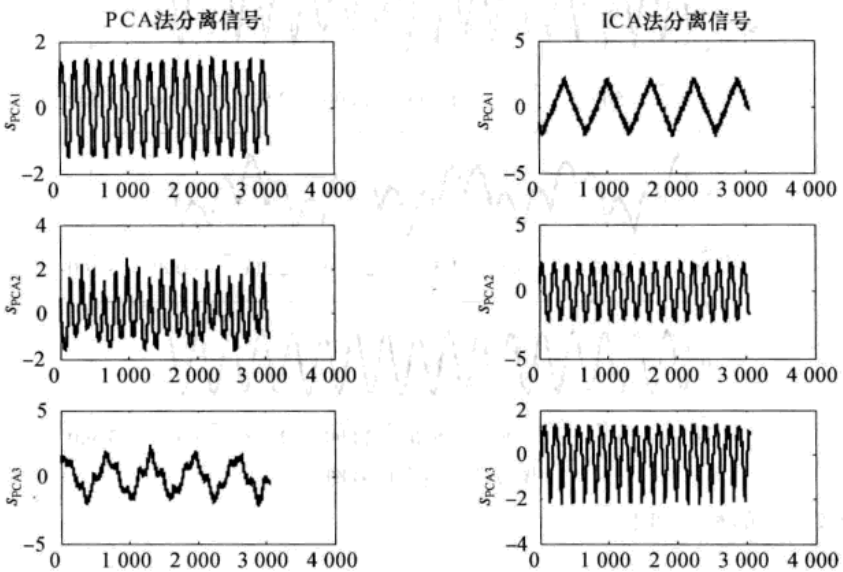


图 5-6　分离信号图形

5.3　典型相关分析

在对经济和管理问题的研究中，不仅需要经常考察两个变量之间的相关程度，而且还需要经常考察多个变量与多个变量之间即两组变量之间的相关性。比如工厂管理人员需要了解原料的主要质量指标 $X_1, X_2, \cdots, X_p$ 与产品的主要质量指标 $Y_1, Y_2, \cdots Y_q$ 之间的相关性，以便提高产品质量；医生要根据病人的一组体检化验指标与一些疾病之间的相关性，确定治疗方法等。典型相关分析就是测度两组变量之间相关程度的一种多元统计方法，它是两个随机变量之间的相关性在两组变量之下的推广。

5.3.1　典型相关分析的基本原理

对于两组随机变量（$X_1, X_2, \cdots, X_p$）和（$Y_1, Y_2, \cdots, Y_q$），像主成分分析那样，考虑（X_1，X_2，…，X_p）的一个线性组合 U 及（$Y_1, Y_2, \cdots, Y_q$）的一个线性组合 V，希望找到的 U 和 V 之间有最大可能的相关系数，以充分反映两组变量间的关系。这样就把研究两组随机变量间相关关系的问题转化为研究两个随机变量间的相关关系。如果一对变量（U，V）还不能完全刻划两组变量间的相关关系，可以继续找第二对变量，希望这对变量在与第一对变量（U，V）不相关的情况下也具有尽可能大的相关系数。直到进行到找不到相关变量对时为止。这便引出了典型相关变量的概念。

1. 总体典型相关变量

设有两组随机向量 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$，$Y=(Y_1,Y_2,\cdots,Y_q)^{\mathrm{T}}(p\leqslant q)$，将两组合并成一组向量 $(X^{\mathrm{T}},Y^{\mathrm{T}})=(X_1,X_2,\cdots,X_p,Y_1,Y_2,\cdots,Y_q)^{\mathrm{T}}$，其协方差矩阵为

$$\Sigma=\begin{bmatrix}\Sigma_{11}\Sigma_{12}\\ \Sigma_{21}\Sigma_{22}\end{bmatrix} \tag{5.3.1}$$

其中 $\Sigma_{11}=\mathrm{cov}(X)$，$\Sigma_{22}=\mathrm{cov}(Y)$，$\Sigma_{12}=\Sigma_{21}^{\mathrm{T}}=\mathrm{cov}(X,Y)$。

根据典型相关思想，问题是要寻找 $X=(X_1,X_2,\cdots,X_p)^{\mathrm{T}}$，$Y=(Y_1,Y_2,\cdots,Y_q)^{\mathrm{T}}(p\leqslant q)$ 的线性组合

$$U_1=a_1^{\mathrm{T}}X=a_{11}X_1+a_{12}X_2+\cdots+a_{1p}X_p$$
$$V_1=b_1^{\mathrm{T}}Y=b_{11}Y_1+b_{12}Y_2+\cdots+b_{1q}Y_q$$

使 U_1、V_1 的相关系数 $\rho(U_1,V_1)$ 达到最大，这里

$$a_1^{\mathrm{T}}=(a_{11},a_{12},\cdots,a_{1p}),b_1^{\mathrm{T}}=(b_{11},b_{12},\cdots,b_{1q})$$

由 (5.3.1) 式，$\mathrm{var}(U_1)=a_1^{\mathrm{T}}\Sigma_{11}a_1$，$\mathrm{var}(V_1)=b_1^{\mathrm{T}}\Sigma_{22}b_1$，$\mathrm{cov}(U_1,V_1)=a_1^{\mathrm{T}}\Sigma_{12}b_1$，所以 U_1、V_1 的相关系数为

$$\rho_{U_1,V_1}=\frac{a_1^{\mathrm{T}}\Sigma_{12}b_1}{\sqrt{a_1^{\mathrm{T}}\Sigma_{11}a_1}\sqrt{b_1^{\mathrm{T}}\Sigma_{22}b_1}} \tag{5.3.2}$$

又由于相关系数与量纲无关，因此可设约束条件

$$a_1^{\mathrm{T}}\Sigma_{11}a_1=b_1^{\mathrm{T}}\Sigma_{22}b_1=1 \tag{5.3.3}$$

满足约束条件 (5.3.3) 的相关系数 $\rho(U_1,V_1)$ 的最大值称为第一典型相关系数，U_1、V_1 称为第一对典型相关变量。

如果 U_1、V_1 还不足以反映 X、Y 之间的相关性，还可构造第二对线性组合：

$$U_2=a_2^{\mathrm{T}}X=a_{21}X_1+a_{22}X_2+\cdots+a_{2p}X_p$$
$$V_2=b_2^{\mathrm{T}}Y=b_{21}Y_1+b_{22}Y_2+\cdots+b_{2q}Y_q$$

使得 (U_1,V_1) 与 (U_2,V_2) 不相关，即

$$\mathrm{cov}(U_1,U_2)=\mathrm{cov}(U_1,V_2)=\mathrm{cov}(U_2,V_1)=\mathrm{cov}(V_1,V_2)=0$$

在约束条件 $\mathrm{var}(U_1)=\mathrm{var}(V_1)=\mathrm{var}(U_2)=\mathrm{var}(V_2)=1$ 下，求 a_2、b_2 使得

$$\rho_{U_2,V_2}=a_2^{\mathrm{T}}\Sigma_{12}b_2$$

取得最大值，此时称 $\rho_{U_2,V_2}=a_2^{\mathrm{T}}\Sigma_{12}b_2$ 为第二典型相关系数，U_2、V_2 为第二对典型相关变量。

一般地，若前 $k-1$ 对典型变量还不足以反映 X、Y 之间的相关性，还可构造第 k 对线性组合：

$$U_k=a_k^{\mathrm{T}}X=a_{k1}X_1+a_{k2}X_2+\cdots+a_{kp}X_p$$
$$V_k=b_k^{\mathrm{T}}Y=b_{k1}Y_1+b_{k2}Y_2+\cdots+b_{kq}Y_q$$

在约束条件

$$\mathrm{var}(U_k)=\mathrm{var}(V_k)=1$$
$$\mathrm{cov}(U_k,U_j)=\mathrm{cov}(U_k,V_j)=\mathrm{cov}(V_k,U_j)=\mathrm{cov}(V_k,V_j)=0\quad(1\leqslant j<k)$$

下，求 a_k、b_k 使得 $\rho_{U_k,V_k}=a_k^{\mathrm{T}}\Sigma_{12}b_k$ 取得最大值。如此确定的 (U_k,V_k) 称为 X、Y 的第 k 对典型变量，相应的 ρ_{U_k,V_k} 称为第 k 个典型相关系数。

2. 总体典型变量与典型相关系数的计算方法

(1) 计算矩阵 $[X^{\mathrm{T}},Y^{\mathrm{T}}]^{\mathrm{T}}$ 的协方差矩阵或相关系数矩阵

$$\Sigma=\begin{bmatrix}\Sigma_{11}\Sigma_{12}\\ \Sigma_{21}\Sigma_{22}\end{bmatrix},R=\begin{bmatrix}R_{11}R_{12}\\ R_{21}R_{22}\end{bmatrix}$$

(2) 令

$$A=(\Sigma_{11})^{-1/2}\Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21}(\Sigma_{11})^{-1/2},B=(\Sigma_{22})^{-1/2}\Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}(\Sigma_{22})^{-1/2}$$

或

$$A=(R_{11})^{-1/2}R_{12}(R_{22})^{-1}R_{21}(R_{11})^{-1/2},B=(R_{22})^{-1/2}R_{21}(R_{11})^{-1}R_{12}(R_{22})^{-1/2}$$

求 A，B 的特征值 $\rho_1^2,\rho_2^2,\cdots,\rho_p^2$ 以及对应的正交单位特征向量 $e_k,f_k,k=1,2,\cdots,p$。

(3) X、Y 的第 k 对典型相关变量为

$$\begin{cases}U_k=a_k^{\mathrm{T}}X=e_k^{\mathrm{T}}\Sigma_{11}^{-0.5}X\\V_k=b_k^{\mathrm{T}}X=f_k^{\mathrm{T}}\Sigma_{22}^{-0.5}Y\end{cases}k=1,2,\cdots,p$$

其中 $\Sigma_{11}^{-0.5}$、$\Sigma_{22}^{-0.5}$ 分别为 Σ_{11}、Σ_{22} 的平方根矩阵的逆矩阵。

(4) X、Y 的第 k 对典型相关变量的相关系数为

$$\rho_k=a_k^{\mathrm{T}}\Sigma_{12}b_k(k=1,2,\cdots,p)$$

以上过程的 MATLAB 实现程序如下：

```
%输入协方差矩阵
X=[data];                          %输入协方差矩阵 X
p=c1;                              %c1 表示 X 向量的维数
q=c2;                              %c2 表示 Y 向量的维数
R11=X(1:p,1:p);                    %读取 Σ11
R12=X(1:p,p+1:p+q);                %读取 Σ12
R21=X(p+1:p+q,1:p);                %读取 Σ21
R22=X(p+1:p+q,p+1:p+q);            %读取 Σ22
[v1,d1]=eig(R11);                  %计算 R11 的特征值与单位正交向量
[v2,d2]=eig(R22);                  %计算 R22 的特征值与单位正交向量
p1=inv(v1*sqrt(d1)*v1');
p2=inv(v2*sqrt(d2)*v2');           %p1,p2 表示 Σ11,Σ22 的平方根矩阵的逆 Σ11^-0.5,Σ22^-0.5
A=p1*R12*inv(R22)*R21*p1;          %计算矩阵 A
B=p2*R21*inv(R11)*R12*p2;          %计算矩阵 B
[va,da]=eig(A),                    %计算 A 的特征值与特征向量
[vb,db]=eig(B),                    %计算 B 的特征值与特征向量
A1=p1*va,                          %计算典型相关变量 U 的系数
B1=p2*vb,                          %计算典型相关变量 V 的系数
r=sqrt(sum(da)),                   %计算典型相关系数
```

根据程序的输出结果写出典型相关变量与典型相关系数。

例 5.3.1 设样本的相关系数矩阵为

$$R=\begin{pmatrix}R_{11}&R_{12}\\R_{21}&R_{22}\end{pmatrix}=\left(\begin{array}{cc:cc}1&0.505&0.569&0.602\\0.505&1&0.422&0.467\\\hdashline0.569&0.422&1&0.926\\0.602&0.467&0.926&1\end{array}\right)$$

计算典型相关系数与典型相关变量。

解：已知相关系数矩阵 R，且 $p=2$，$q=2$，计算程序如下：

```
X=[1,0.505,0.569,0.602;0.505,1,0.422,0.467;0.569,0.422,1,0.926;0.602,0.467,0.926,1];
p=2;
q=2;
R11=X(1:p,1:p);
```

```
R12= X(1:p,p+ 1:p+ q);
R21= X(p+ 1:p+ q,1:p);
R22= X(p+ 1:p+ q,p+ 1:p+ q);
[v1,d1]= eig(R11);                    %计算 R11 的特征值与单位正交向量
[v2,d2]= eig(R22);                    %计算 R22 的特征值与单位正交向量
p1= inv(v1*sqrt(d1)*v1');
p2= inv(v2*sqrt(d2)*v2');
A= p1*R12*inv(R22)*R21*p1;            %计算矩阵 A
B= p2*R21*inv(R11)*R12*p2;            %计算矩阵 B
 [va,da]= eig(A);
[vb,db]= eig(B);
A1= p1*va                             %计算典型相关变量 U 的系数
B1= p2*vb                             %计算典型相关变量 V 的系数
r= sqrt(sum(da))                      %计算典型相关系数
```

输出结果为：

```
A1 =
    0.7808   - 0.8560
    0.3445     1.1062
B1 =
   - 2.6482   - 0.0603
     2.4749   - 0.9439
r =
   0.6311    0.0568
```

所以典型变量为：

$$U_1 = 0.780\,8X_1 + 0.344\,5X_2$$
$$V_1 = 0.060\,3Y_1 + 0.943\,9Y_2$$
$$U_2 = -0.856\,0X_1 + 1.106\,2X_2$$
$$V_2 = -2.648\,2Y_1 + 2.474\,9Y_2$$

典型相关系数为：$\rho_1 = 0.6311$，$\rho_2 = 0.0568$。

5.3.2　样本的典型变量与典型相关系数

在实际问题中，$(X^{\mathrm{T}}, Y^{\mathrm{T}})^{\mathrm{T}}$ 的协方差矩阵 Σ（或相关系数矩阵 R）一般是未知的，我们所具有的资料通常是关于 X 和 Y 的 n 组观测数据：

$$X_i = (x_{i1}, x_{i2}, \cdots, x_{in})^{\mathrm{T}} \quad (i = 1,2,\cdots,p)$$
$$Y_j = (y_{j1}, y_{j2}, \cdots, y_{jn})^{\mathrm{T}} \quad (j = 1,2,\cdots,q)$$

同主成分分析一样，将这些观测数据的样本协方差矩阵

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \text{或} R = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

作为 Σ 或 ρ 的估计，其中 $S_{11} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})^{\mathrm{T}}$，$S_{22} = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})(Y_i - \overline{Y})^{\mathrm{T}}$，

$S_{12} = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})^{\mathrm{T}}$，$S_{21} = S_{12}^{\mathrm{T}}$

以 S 代替 Σ 或 R 代替 ρ 所求得的典型变量和典型相关系数分别称为样本典型变量和样本

典型相关系数。以 S 代替 Σ 或 R 代替 ρ，此时样本典型变量和典型相关系数计算方法同总体典型变量和典型相关系数的计算方法一样。

在 MATLAB 中，样本典型相关分析的命令为 canoncorr，其调用格式为：

```
[A,B,r,U,V,stats] = canoncorr(X,Y)
```

其中输入 X 表示第一组向量的观测矩阵，Y 表示第二组向量的观测矩阵；输出 A、B 是典型相关变量的系数矩阵；r 表示典型相关系数；U、V 表示典型相关变量的得分；输出 stats 包括 wilks、chisq 及 f 统计量以及相应的概率。

例 5.3.2 某康复俱乐部对 20 名中年人测量了三项生理指标：体重（weight）、腰围（waist）、脉搏（pulse）和三项训练指标：引体向上（chins）、起坐次数（situps）、跳跃次数（jumps）。其数据列于表 5-6。试分析这两组变量间的相关性。

表 5-6 某康复俱乐部测量的 20 名中年人的生理指标和训练指标

Obs	weight	waist	pulse	chins	situps	jumps	Obs	weight	waist	pulse	chins	situps	jumps
1	191	36	50	5	162	60	11	169	34	50	17	120	38
2	189	37	52	2	110	60	12	166	33	52	13	210	115
3	193	38	58	12	101	101	13	154	34	64	14	215	105
4	162	35	62	12	105	37	14	247	46	50	1	50	50
5	189	35	46	13	155	58	15	193	36	46	6	70	31
6	182	36	56	4	101	42	16	202	37	62	12	210	120
7	211	38	56	8	101	38	17	176	37	54	4	60	25
8	167	34	60	6	125	40	18	157	32	52	11	230	80
9	176	31	74	15	200	40	19	156	33	54	15	225	73
10	154	33	56	17	251	250	20	138	33	68	2	110	43

解：三项生理指标作为第一组向量 X，三项训练指标作为第二组向量 Y，表 5-6 中的数据作为样本数据，调用典型相关分析命令。程序如下：

```
DATA= [⋯];                                  %将表 5-6 中的数据输入 DATA
X= DATA(:,1:3);                             %第一组向量观测值
Y= DATA(:,4:6);                             %第二组向量观测值
[A,B,r,U,V,stats] = canoncorr(X,Y);
A,B,r
```

输出结果为：

```
A =
   - 0.0314   - 0.0763    0.0077
     0.4932     0.3687  - 0.1580
   - 0.0082   - 0.0321  - 0.1457
B =
   - 0.0661   - 0.0710    0.2453
   - 0.0168     0.0020  - 0.0198
     0.0140     0.0207    0.0082
r =
   0.7956    0.2006    0.0726
```

5.3.3 典型相关系数的显著性检验

典型相关分析是否恰当，应该取决于两组原变量之间是否相关，如果两组变量之间毫无

相关性而言，则不应该作典型相关分析。用样本来估计总体的典型相关系数是否有误，需要进行检验。

1. 检验方法

设总体 X，Y 的各对典型相关系数为 $\rho_1 \geqslant \rho_2 \geqslant \cdots \rho_p \geqslant 0$，首先提出检验的原假设与备择假设

$$H_0^{(1)}:\rho_1 = 0 \leftrightarrow H_1^{(1)}:\rho_1 \neq 0$$

若不能拒绝原假设，则 $\rho_1 = \rho_2 = \cdots = \rho_p = 0$，此时不能做典型相关分析；若拒绝 $H_0^{(1)}$，继续如下检验

$$H_0^{(2)}:\rho_2 = 0 \leftrightarrow H_1^{(2)}:\rho_2 \neq 0$$

若不能拒绝 $H_0^{(2)}$，表明只有第一对典型变量显著相关外，其余变量均不显著，实际应用时只需要考虑第一对典型变量；若拒绝 $H_0^{(2)}$，则需检验 ρ_3 是否为零，以此类推，若假设 $\rho_{k-1}=0$被拒绝，则检验

$$H_0^{(k)}:\rho_k = 0 \leftrightarrow H_1^{(k)}:\rho_k \neq 0$$

若不能拒绝 $H_0^{(k)}$，则只需考虑前 $k-1$ 对典型相关变量，否则继续检验，直至检验 ρ_p 是否为零。

在总体服从 $p+q$ 维正态分布条件下，可用如下的似然比统计量进行检验

$$T_k = -[n-(p+q+3)/2]\ln\Lambda_k \sim \chi^2(d_{1k})$$

其中 $\Lambda_k = \prod_{j=k}^{p}(1-\hat{\rho}_j^2)$，$d_{1k} = (p-k+1)(q-k+1)$，对于给定的 α，计算概率

$$p_k = P_{H_0}(T_k \geqslant t_k) = P(\chi^2(d_{1k}) \geqslant t_k)$$

若 $\rho_k < \alpha$，即认为第 k 对典型变量显著相关。上述检验依次对 $k=1,2,\cdots,p$ 进行，若对某个 k 检验概率首次大于 α，则检验停止，即认为只有前 $k-1$ 对典型变量显著相关。

2. 典型相关分析检验的 MATLAB 实现

设 $X=(x_{ij})_{n\times p}$，$Y=(y_{ik})_{n\times q}$ 是取自总体的观测数据，利用 MATLAB 软件进行典型相关分析的步骤如下：

1）输入数据并计算协方差矩阵或相关系数矩阵：

```
a= [X,Y];          %此前 X,Y 的数据应该已经输入
[n,m]= size(a);
S= cov(a);         %协方差矩阵
R= corref(a);      %相关系数矩阵
```

2）计算典型相关系数：

```
R1= inv(R(1:p,1:p)) *R(1:p,p+ 1:p+ q) *inv(R(p+ 1:p+ q, p+ 1:p+ q)) *R(p+ 1:p+ q,1:p);
d= sort(eig(R1),'descend');
xgxs= sqrt(d);
```

3）计算典型相关向量：

```
X= X./[ones(n,1) *std(X)];
Y= Y./[ones(n,1) *std(Y)];
[A,B] = canoncorr(X,Y);
U= (X- ones(n,1) *mean(X)) *A
V= (Y- ones(n,1) *mean(Y)) *B
```

4）典型相关系数的显著性检验：

```
D= 1- d;
```

```
f1= fliplr(D');                          %矩阵左右翻转
f2= cumprod(f1);                         %向量累积乘积
k= 1:p;
dlk= (p- k+ 1).*(q- k+ 1);
Qk= - [n- 0.5*(p+ q+ 3)].*(log(fliplr(f2)));
GL= 1- chi2cdf(Qk,dlk);
```

5.3.4　典型相关分析实例

例 5.3.3　选取 1980—2008 年安徽省人均粮食总产量（吨/人）、人均农业总产值（亿元/万人）、人均粮食播种面积（千公顷/万人）、人均农业机械总动力（千瓦/人）、单位面积化肥施用（万吨/千公顷）、人均受灾面积（千公顷/万人）以及农业生产资料价格指数指标，分别记为：x_1，x_2，x_3，y_1，y_2，y_3，y_4，见表 5-7。解决以下问题：（1）对安徽省粮食生产进行主成分分析，在此基础上给出适当的分类；（2）对安徽省粮食生产影响因素进行典型相关分析。

表 5-7　1980—2008 年安徽省粮食产出及影响因素数据

年　份	x_1	x_2	x_3	y_1	y_2	y_3	y_4
1980	0.870 4	0.041 1	4.633 2	0.397 9	0.007 1	0.262 8	102.100 0
1981	1.053 8	0.056 8	4.566 4	0.392 9	0.009 1	0.613 0	101.700 0
1982	1.081 8	0.058 6	4.480 8	0.404 7	0.011 4	0.234 9	101.300 0
1983	1.089 8	0.060 4	4.260 0	0.414 7	0.011 5	0.133 9	102.800 0
1984	1.157 6	0.066 4	4.187 2	0.419 1	0.012 7	0.403 6	107.000 0
1985	1.098 3	0.073 6	4.147 0	0.422 3	0.013 9	0.213 1	101.700 0
1986	1.164 9	0.081 7	4.008 9	0.450 3	0.014 1	0.361 7	102.100 0
1987	1.167 0	0.090 2	4.022 6	0.497 8	0.014 4	0.289 5	112.800 0
1988	1.066 1	0.099 2	3.769 6	0.529 7	0.015 5	0.689 4	118.600 0
1989	1.088 0	0.106 0	3.696 9	0.549 2	0.016 7	0.333 3	121.700 0
1990	1.095 0	0.113 4	3.612 3	0.568 0	0.017 4	0.482 1	103.900 0
1991	0.741 0	0.087 4	3.472 0	0.584 7	0.017 6	0.457 6	102.300 0
1992	0.962 8	0.107 8	3.352 7	0.597 0	0.019 1	0.375 4	102.500 0
1993	1.037 4	0.142 7	3.303 0	0.620 3	0.021 5	0.571 7	112.900 0
1994	0.928 6	0.199 5	3.249 9	0.662 1	0.023 0	0.162 6	122.800 0
1995	1.023 3	0.246 1	3.222 8	0.708 3	0.024 3	0.363 7	128.000 0
1996	1.031 2	0.261 1	3.193 0	0.770 2	0.029 7	0.226 1	107.200 0
1997	1.047 8	0.262 1	3.155 0	0.837 3	0.028 5	0.224 7	98.900 0
1998	0.953 3	0.250 1	3.151 5	0.937 1	0.029 6	0.142 4	94.800 0
1999	1.017 2	0.259 4	3.150 2	1.015 4	0.029 8	0.333 6	95.300 0
2000	0.883 6	0.241 4	3.008 8	1.063 6	0.030 1	0.347 0	98.200 0
2001	0.886 2	0.243 8	2.919 1	1.121 7	0.031 8	0.262 8	97.900 0
2002	0.973 1	0.250 7	2.958 0	1.186 8	0.032 2	0.613 0	99.900 0
2003	0.773 9	0.215 9	2.943 8	1.238 7	0.033 4	0.234 9	100.200 0
2004	0.942 4	0.289 3	2.965 6	1.300 1	0.032 2	0.133 9	112.000 0
2005	0.886 4	0.278 5	2.978 8	1.355 4	0.032 6	0.403 6	108.300 0
2006	0.958 4	0.304 0	2.948 7	1.420 4	0.033 4	0.213 1	100.000 0
2007	0.967 8	0.351 6	2.953 4	1.512 8	0.034 5	0.361 7	106.800 0
2008	0.996 9	0.395 0	2.957 0	1.585 2	0.034 3	0.289 5	123.900 0

资料来源：1980—2004 年数据由《安徽五十年》有关数据整理得到，2005—2008 年数据来源于《安徽统计年鉴》2006—2009。

解：(1) 设原始数据矩阵为

$$a = (a_{ij})_{29\times 7}$$

对 a 进行标准化的无量纲变换，得到矩阵

$$b = (b_{ij})_{29\times 7}$$

其中 $b_{ij} = a_{ij}/s_j, s_j = \sqrt{\frac{1}{28}\sum_{i=1}^{29}(a_{ij}-\overline{a}_j)^2}, \overline{a}_j = \frac{1}{29}\sum_{i=1}^{29}a_{ij}, j = 1,2,\cdots,7$。

由于原始数据的协方差矩阵与相关系数矩阵得到的最大特征值对应的特征向量不是正向量，所以我们采用 R 矩阵进行主成分分析。由 R 矩阵的定义

$$R = (r_{ij})_{7\times 7}, r_{ij} = \frac{2\sum_{k=1}^{29} b_{ki}b_{kj}}{\sum_{k=1}^{29} b_{ki}^2 + \sum_{k=1}^{29} b_{kj}^2}$$

实对称矩阵 R 的特征值与对应的特征向量及贡献率见表 5-8。

表 5-8　特征值、特征向量及贡献率

特 征 值	特 征 向 量	贡 献 率	累积贡献率
4.832 9	(0.348 9，0.381 4，0.385 1，0.404 4，0.419 5，0.381 3，0.315 5)	0.690 4	0.690 4
1.662 9	(0.492 3，−0.372 8，0.345 2，−0.332 6，−0.268 1，−0.209 5，0.524 4)	0.237 6	0.938 0
0.313 2	(0.088 0，0.390 0，−0.248 8，0.224 5，0.132 6，−0.806 3，0.245 3)	0.044 7	0.972 7
0.133 0	(−0.022 3，−0.097 8，0.705 0，0.027 5，0.200 9，−0.392 6，−0.545 6)	0.019 0	0.991 7
0.034 8	(−0.124 6，−0.661 2，−0.192 8，0.142 0，0.673 0，−0.069 2，0.179 3)	0.005 0	0.996 7
0.014 0	(0.134 5，0.310 1，−0.102 5，−0.791 9，0.491 1，0.033 3，−0.076 7)	0.002 0	0.998 7
0.009 2	(0.770 8，−0.142 9，−0.357 5，0.165 7，−0.004 8，0.024 9，−0.479 2)	0.001 3	1

由于第一、第二主成分累积贡献率达到 93.8%，故选择两个主成分计算主成分得分：

$F_1 = 0.348\,9b_1 + 0.381\,4b_2 + 0.385\,1b_3 + 0.404\,4b_4 + 0.419\,5b_5 + 0.381\,3b_6 + 0.315\,5b_7$

$F_2 = 0.492\,3b_1 - 0.378\,2b_2 + 0.345\,2b_3 - 0.332\,6b_4 - 0.268\,1b_5 - 0.209\,5b_6 + 0.524\,4b_7$

根据主成分得分图（图 5-7）知：安徽省农业生产分为三阶段：1980—1987，1988—1995，1996—2008。

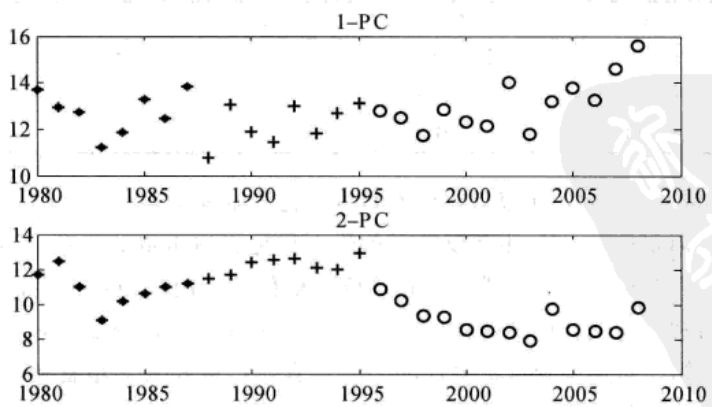


图 5-7　安徽省农业生产主成分得分图

(2) 为了分析影响安徽省粮食生产的因素，令 $X=(b_1, b_2, b_3)$，$Y=(b_4, b_5, b_6, b_7)$ 首

先计算［X，Y］的协方差矩阵

$$\Sigma=\begin{pmatrix}\Sigma_{11}\Sigma_{12}\\ \Sigma_{21}\Sigma_{22}\end{pmatrix}=\begin{bmatrix}1.000\,0 & -0.358\,1 & 0.500\,2 & -0.485\,9 & -0.472\,2 & 0.111\,4 & 0.200\,2\\ -0.358\,1 & 1.000\,0 & -0.876\,9 & 0.926\,5 & 0.952\,7 & -0.232\,0 & 0.100\,0\\ 0.500\,2 & -0.876\,9 & 1.000\,0 & -0.837\,0 & -0.952\,3 & 0.103\,4 & -0.054\,5\\ -0.485\,9 & 0.926\,5 & -0.837\,0 & 1.000\,0 & 0.931\,8 & -0.167\,0 & -0.027\,4\\ -0.472\,2 & 0.952\,7 & -0.952\,3 & 0.931\,8 & 1.000\,0 & -0.202\,9 & -0.045\,6\\ 0.111\,4 & -0.232\,0 & 0.103\,4 & -0.167\,0 & -0.202\,9 & 1.000\,0 & 0.115\,9\\ 0.200\,2 & 0.100\,0 & -0.054\,5 & -0.027\,4 & -0.045\,6 & 0.115\,9 & 1.000\,0\end{bmatrix}$$

其次令

$$A=(\Sigma_{11})^{-1/2}\Sigma_{12}(\Sigma_{22})^{-1}\Sigma_{21}(\Sigma_{11})^{-1/2},B=(\Sigma_{22})^{-1/2}\Sigma_{21}(\Sigma_{11})^{-1}\Sigma_{12}(\Sigma_{22})^{-1/2}$$

求 A，B 的特征值 ρ_1^2，ρ_2^2，ρ_3^2 以及对应的正交单位特征向量 e_k，$f_k(k=1,2,3)$，得到 X，Y 的 3 对典型相关变量为

$$\begin{aligned}U_1&=0.003\,8x_1-0.510\,2x_2+0.520\,3x_3\\ V_1&=0.064\,6y_1-1.060\,9y_2-0.016\,0y_3-0.124\,0y_4\\ U_2&=0.643\,2x_1-1.912\,6x_2-2.205\,5x_3\\ V_2&=-2.495\,0y_1+2.388\,5y_2+0.484\,2y_3+0.069\,4y_4\\ U_3&=-0.985\,8x_1-0.760\,0x_2-0.304\,3x_3\\ V_3&=0.080\,7y_1+0.051\,6y_2+0.269\,8y_3-0.987\,9y_4\end{aligned}$$

典型相关系数为

$$\rho_1=0.991\,5,\quad \rho_2=0.677\,1,\quad \rho_3=0.267\,1$$

最后对典型相关系数进行检验（$\alpha=0.05$）

$$\begin{aligned}H_0^{(1)}:\rho_1=0\leftrightarrow H_1^{(1)}:\rho_1\neq 0\\ H_0^{(2)}:\rho_2=0\leftrightarrow H_1^{(2)}:\rho_2\neq 0\\ H_0^{(3)}:\rho_3=0\leftrightarrow H_1^{(3)}:\rho_3\neq 0\end{aligned}$$

检验结果见表 5-9。

表 5-9　各对典型变量相关的显著性检验

k	Λ_k	F_k	d_{1k}	d_{2k}	p_k
1	0.008 5	24.648 5	12	58.498 0	0
2	0.502 9	3.144 4	6	46	0.011 3
3	0.928 7	0.921 8	2	24	0.393 6

因为 $p_3=0.393\,6>0.05$，所以只有前两对典型变量显著相关，由于 U_1 主要反映了人均农业产值与人均耕地面积信息，V_1 主要反映了单位面积化肥施用量的信息，因此第一对典型变量主要反映了人均农业产值与人均耕地面和单位面积化肥施用量的相关性；同理，U_2 主要提取人均农业总产值和人均粮食播种面积信息，V_2 主要反映了人均农业机械总动力和单位面积化肥施用。因此，第二对典型变量主要反映了人均农业产值和人均粮食播种面积与人均农业机械总动力、单位面积化肥施用量的相关性。

实例的 MATLAB 程序如下：

```
%首先输入原始数据(表 5-7 中数据作为 data 放入矩阵 a)
a=[data];
```

```
[n,m]= size(a);
%主成分分析程序
b= a./(ones(n,1)*std(a));
for i= 1:m,
      for j= 1:m
      C(i,j)= [2*dot(b(:,i),b(:,j))]./[sum(b(:,i).^2)+ sum(b(:,j).^2)];
      end
 end
 [v,d]= eig(C);
 F= b*v;
%分类程序
[c1,u1,obj_fcn]= fcm(b,3);
f1= sort(u1);
[f1,I1]= sort(u1);
d1= find(I1(3,:)= = 1);
d2= find(I1(3,:)= = 2);
d3= find(I1(3,:)= = 3);
subplot(211),plot(1980:1987,F(d1,7),'*'),
hold on,
plot(1988:1995,F(d2,7),'+ '),
hold on,
plot(1996:2008,F(d3,7),'or'),title('1- PC')
subplot(212),
plot(1980:1987,F(d1,6),'*'),
hold on,
plot(1988:1995,F(d2,6),'+ '),
hold on,
plot(1996:2008,F(d3,6),'or'),title('2- PC')
%典型相关分析程序
R= cov(b);
p= 3;q= m- p;
X= b(:,1:p);
Y= b(:,p+ 1:m);
[A,B,r,U,V,stats]= canoncorr(X,Y);
```

习　题　5

1. 设随机向量 $X=(X_1, X_2, X_3)^{\mathrm{T}}$ 的协方差与相关系数矩阵分别为

$$\Sigma=\begin{pmatrix}1 & 4\\ 4 & 25\end{pmatrix},\quad R=\begin{pmatrix}1 & 0.8\\ 0.8 & 1\end{pmatrix}$$

分别从 Σ，R 出发，求 X 的各主成分以及各主成分的贡献率并比较差异情况。

2. 设 $A=[a_1 \quad a_2 \quad \cdots \quad a_6]=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{16}\\ a_{21} & a_{22} & \cdots & a_{26}\\ \vdots & \vdots & & \vdots\\ a_{51} & a_{52} & \cdots & a_{56}\end{bmatrix}$，表示某球队参加 5 场比赛的技术统计，其中 a_1 表

示一攻得分，a_2 表示快攻得分，a_3 表示拦网得分，a_4 表示失误次数，a_5 表示发球得分，a_6 表示一传到位率。根据以上资料进行主成分分析时，请回答以下问题：

(1) 原始数据是否需要进行处理？若需处理写出你处理的方法（计算公式）。

(2) 利用协方差矩阵和相关系数矩阵计算主成分得分时有何不同？

(3) 进行主成分分析时，特征值的作用是什么？特征向量的意义是什么？

(4) 选择主成分个数的依据是什么？

(5) 如下的两个主成分主要反映哪些指标的作用？给主成分起一个恰当的指标名称。

$$F_1 = 0.477\,3a_1 + 0.636\,4a_2 - 0.079\,6a_3 - 0.159\,1a_4 + 0.556\,9a_5 + 0.159\,1a_6$$

$$F_2 = 0.150\,4a_1 + 0.150\,4a_2 + 0.601\,5a_3 + 0.157\,01a_4 - 0.075\,2a_5 + 0.526\,41a_6$$

3. 根据安徽省各地市经济指标数据，见表 5-10，解决以下问题：

(1) 利用主成分分析对 2007 年安徽省 17 个地市的经济发展进行分析，给出排名。

(2) 此时能否只用第一主成分进行排名？为什么？

表 5-10 安徽省各市“三资”工业企业主要经济指标（2007 年）

地　区	工业总产值	资产合计	工业增加值	实收资本	长期负债	业务收入	业务成本	利　润
合肥	491.70	380.31	158.39	121.54	22.74	439.65	344.44	17.43
淮北	21.12	30.55	6.40	12.40	3.31	21.17	17.71	2.03
亳州	1.71	2.35	0.57	0.68	0.13	1.48	1.36	−0.03
宿州	9.83	9.05	3.13	3.43	0.64	8.76	7.81	0.54
蚌埠	64.06	77.86	20.63	30.37	5.96	63.57	52.15	4.71
阜阳	30.38	46.90	9.19	9.83	17.87	28.24	21.90	3.80
淮南	31.20	70.07	8.93	18.88	33.05	31.17	26.50	2.84
滁州	79.18	62.09	20.78	24.47	3.51	71.29	59.07	6.78
六安	47.81	40.14	17.50	9.52	4.14	45.70	34.73	4.47
马鞍山	104.69	78.95	29.61	25.96	5.39	98.08	84.81	3.81
巢湖	21.07	17.83	6.21	6.22	1.90	20.24	16.46	1.09
芜湖	214.19	146.78	65.16	41.62	4.39	194.98	171.98	11.05
宣城	31.16	27.56	8.80	9.44	1.47	28.83	25.22	1.05
铜陵	12.79	14.16	3.66	4.07	1.57	11.95	10.24	0.73
池州	6.45	5.37	2.39	2.20	0.40	5.97	4.79	0.52
安庆	39.43	44.60	15.17	15.72	3.27	36.03	27.87	3.48
黄山	5.02	3.62	1.63	1.42	0.53	4.45	4.04	0.02

资料来源：《安徽统计年鉴 2008》。

4. 根据 1998 年部分地区洪灾损失数据（见表 5-11），进行主成分分析，哪些省受灾较轻？受灾最重的是哪三个省？其中 $x_1 \sim x_{12}$ 分别为：受灾、成灾、绝收面积、受灾/万人次、成灾/万人次、死亡/人、伤病/人、紧急转移/人、倒塌房屋/万间、损坏房屋/万间、死亡大牲畜/万头、直接经济损失/亿元。

表 5-11 1998 年部分地区洪灾损失指标数据

地　区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
蒙	130.4	107.7	64.8	448.0	375.0	147.0	105 476.0	97.7	37.0	59.0	36.0	164.0
吉	109.7	64.7	26.7	306.1	214.5	7.0	311 000.0	98.3	56.0	63.2	14.8	140.0
黑	242.9	160.6	93.7	581.0	521.0	2.0	316 844.0	156.4	82.0	75.0	16.1	218.0
皖	199.6	130.8	57.1	1 562.0	1 012.8	93.0	262 461.0	100.4	26.9	49.3	1.0	130.5

（续）

地　区	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
闽	69.7	23.7	4.4	597.5	455.8	146.0	8 402.0	195.3	68.9	167.8	100.0	87.9
赣	241.6	193.6	87.8	2 381.0	1 702.6	237.0	126 505.0	304.6	117.9	168.2	50.7	434.2
鄂	254.0	169.0	44.5	1 939.0	1 534.7	353.0	891 200.0	247.4	86.9	241.2	27.0	357.0
湘	213.0	141.3	39.9	2 178.0	1 652.4	854.0	224 400.0	350.8	93.9	224.5	83.9	422.8
贵	79.3	54.8	28.2	1 378.5	992.8	315.0	38 000.0	81.1	10.7	76.3	6.6	114.9
川	128.2	75.4	16.9	1 757.9	1 044.3	581.0	14 806.0	37.0	21.2	43.1	13.1	74.7
渝	65.3	49.4	6.7	904.0	668.2	304.0	21 715.0	39.1	17.1	27.1	4.5	55.5
滇	39.9	16.1	3.5	462.7	52.8	166.0	235.0	7.6	10.4	14.8	1.5	23.1
陕	40.8	26.2	4.8	650.0	475.0	215.0	3 069.0	12.7	9.9	24.6	1.8	43.0

资料来源：李琼，周建中．改进主成分分析法在洪灾损失评估中的应用［J］．水电能源科学，2010，28（3）．

5. 已知二维随机向量 $X^{(1)}$，$X^{(2)}$ 的均值向量为（−3，2），（0，1），协方差矩阵为

$$\Sigma=\begin{pmatrix}\Sigma_{11}\Sigma_{12}\\ \Sigma_{21}\Sigma_{22}\end{pmatrix}=\begin{bmatrix}8 & 2 & 3 & 1\\ 2 & 5 & -1 & 3\\ 3 & -1 & 6 & -2\\ 1 & 3 & -2 & 7\end{bmatrix}$$

（1）计算典型相关系数 ρ_1，ρ_2。

（2）确定典型相关变量 (U_1, V_1)，(U_2, V_2)。

实验 4　主成分分析与典型相关分析

实验目的

1. 熟练掌握利用 MATLAB 进行主成分分析的计算步骤。
2. 掌握选择主成分个数的原则以及利用特征值建立权向量的方法。
3. 能根据主成分的数学公式，针对实际问题给出主成分的合理解释。
4. 掌握典型相关分析的方法。

实验数据与内容

1. 主成分分析实验

实验数据见表 5-12。

表 5-12　各地区国有及国有控股工业企业主要经济效益指标（2007 年）

地　区	工业增加值率	总资产贡献率	资产负债率	流动资产周转次数	工业成本费用利润率	产品销售率
北　京	25.92	5.52	34.04	2.05	7.93	99.19
天　津	34.29	16.18	62.66	2.62	12.44	99.58
河　北	29.46	11.87	61.02	2.53	7.23	99.34
山　西	37.58	11.28	67.65	1.95	8.68	98.18
内蒙古	47.36	11.43	62.23	2.21	13.80	99.08

（续）

地　区	工业增加值率	总资产贡献率	资产负债率	流动资产周转次数	工业成本费用利润率	产品销售率
辽　宁	28.73	8.86	60.88	2.17	4.14	99.21
吉　林	30.31	15.14	58.53	2.66	9.26	95.97
黑龙江	52.12	33.67	55.26	2.56	32.94	99.21
上　海	27.39	12.42	45.62	2.13	8.04	99.26
江　苏	26.45	14.02	58.99	2.88	6.91	99.64
浙　江	24.48	14.82	58.81	3.18	6.17	99.65
安　徽	35.13	10.63	65.65	2.39	4.95	98.40
福　建	29.76	12.67	59.34	2.41	8.11	99.54
江　西	26.75	12.00	65.12	2.51	5.60	98.69
山　东	31.60	17.64	59.02	2.94	9.91	99.36
河　南	37.70	13.02	65.02	2.68	6.86	98.58
湖　北	33.75	10.65	54.28	2.17	9.87	98.56
湖　南	35.96	16.62	62.35	2.62	7.00	99.32
广　东	32.84	17.68	48.65	2.88	12.85	99.36
广　西	32.31	12.12	64.04	2.45	7.72	101.20
海　南	35.02	13.59	49.41	2.34	14.50	101.23
重　庆	32.96	11.97	59.24	2.03	5.97	96.58
四　川	37.00	10.72	63.54	1.70	8.62	98.80
贵　州	37.49	12.52	65.69	1.86	8.97	98.35
云　南	41.22	20.94	49.16	1.85	12.44	99.42
西　藏	63.03	3.32	20.40	0.53	10.76	90.38
陕　西	43.67	16.61	57.28	1.91	17.82	98.36
甘　肃	26.57	13.42	58.38	2.54	7.51	98.31
青　海	41.62	14.58	63.15	1.92	26.59	98.11
宁　夏	38.98	8.27	63.53	1.83	5.07	98.53
新　疆	45.58	25.84	49.36	3.16	29.88	100.29

（1）根据指标的属性将原始数据统一趋势化。

（2）利用协方差、相关系数矩阵进行主成分分析，可否只用第一主成分排名？

（3）构造新的实对称矩阵，使得可以只用第一主成分排名。

（4）排名的结果是否合理？为什么？

2. 典型相关分析实验

为研究空气温度与土壤温度的关系，考虑如下 6 个变量：X_1（日最高土壤温度）、X_2（日最低土壤温度）、X_3（日土壤温度曲线积分值）、Y_1（日最高气温）、Y_2（日最低气温）、Y_3（日气温曲线积分值），共观测了 46 天，数据见表 5-13，令 $X=(X_1, X_2, X_3)^{\mathrm{T}}$，$Y=(Y_1, Y_2, Y_3)^{\mathrm{T}}$，对 X、Y 做典型相关分析。

表 5-13　日土壤温度与日气温数据

序　号	X_1	X_2	X_3	Y_1	Y_2	Y_3
1	85	59	151	84	65	147
2	86	61	159	84	65	149
3	83	64	152	79	66	142

（续）

序　号	X_1	X_2	X_3	Y_1	Y_2	Y_3
4	83	65	158	81	67	147
5	88	69	180	84	68	167
6	77	67	147	74	66	131
7	78	69	159	73	66	131
8	84	68	159	75	67	134
9	89	71	195	84	68	161
10	91	76	206	86	72	169
11	91	76	206	88	73	176
12	94	76	211	90	74	187
13	94	75	211	88	72	171
14	92	70	201	58	72	171
15	87	68	167	81	69	154
16	83	68	162	79	68	149
17	87	66	173	84	69	160
18	87	68	177	84	70	160
19	88	70	169	84	70	168
20	83	66	170	77	67	147
21	92	67	196	87	67	166
22	92	72	199	89	69	171
23	94	72	204	89	72	180
24	92	73	201	93	72	186
25	93	72	206	93	74	188
26	94	72	208	94	75	199
27	95	73	214	93	74	193
28	95	70	210	93	74	196
29	95	71	207	96	75	198
30	95	69	202	95	76	202
31	96	69	173	84	73	173
32	91	69	168	91	71	170
33	89	70	189	88	72	179
34	95	71	210	89	72	179
35	96	73	208	91	72	182
36	97	75	215	92	74	196
37	96	69	198	94	75	192
38	95	67	196	96	75	195
39	94	75	211	93	76	198
40	92	73	198	88	74	188
41	90	74	197	88	74	178
42	94	70	205	91	72	175
43	95	71	209	92	72	190
44	96	72	208	92	73	189
45	95	71	208	94	75	194
46	96	71	208	96	76	202

数据来源：梅长林，范金城．数据分析方法［M］．北京：高等教育出版社，2006：134.

第6章 聚类分析

对事物进行分类，是人们认识事物的出发点，也是人们认识世界的一种重要方法。因此，分类学已成为人们认识世界的一门基础学科。聚类分析又称群分析，它是研究（样品或指标）分类问题的一种多元统计方法。所谓类，通俗地说，就是指相似元素的集合。本章主要介绍谱系聚类、K均值聚类、模糊C均值聚类和模糊减法聚类及它们的MATLAB实现。

6.1 距离聚类

6.1.1 聚类的思想

在社会经济领域中存在着大量分类问题，比如对我国多个省、市、自治区独立核算工业企业经济效益进行分析，一般不是逐个省、市、自治区去分析，而较好的做法是选取能反映企业经济效益的代表性指标，如百元固定资产实现利税、资金利税率、产值利税率、百元销售收入实现利润、全员劳动生产率等，根据这些指标对多个省、市、自治区进行分类，然后根据分类结果对企业经济效益进行综合评价，就易于得出科学的分析。又比如，对某些大城市的物价指数进行考察，而物价指数很多，有农用生产物价指数、服务项目物价指数、食品消费物价指数、建材零售价格指数等。由于要考察的物价指数很多，通常先对这些物价指数进行分类。总之，需要分类的问题很多，因此聚类分析这个有用的数学工具越来越受到人们的重视，它在许多领域中都得到了广泛的应用。

聚类问题的一般提法是：设有 n 个样品的 p 元观测数据组成一个数据矩阵

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

其中每一行表示一个样品，每一列表示一个指标，x_{ij} 表示第 i 个样品关于第 j 项指标的观测值，要根据观测值对样品或指标进行分类。一种分类的思想是：在样品之间定义距离，在指标之间定义相似系数。样品距离表明样品之间的相似度，指标之间的相似系数刻画指标之间的相似度。将样品（或变量）按相似度的大小逐一归类，关系密切的聚集到较小的一类，关系疏远的聚集到较大的一类，直到所有的样品（或变量）都聚集完毕。上述思想正是聚类分析的基本思想。

值得注意的是：第4章介绍的判别分析和聚类分析是两种不同目的的分类方法，它们所起的作用是不同的。判别分析方法假定组（或类）已事先分好，判别新样品应归属哪一组，

对组的事先划分有时也可以通过聚类分析得到。聚类分析方法是按样品（或变量）的数据特征，倾向于把相似的样品（或变量）分在同一类中，把不相似的样品（或变量）分在不同类中。

6.1.2 向量的距离

设有 n 个样品的 p 元观测数据

$$x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}} \quad (i = 1, 2, \cdots, n)$$

这时，每个样品可看成 p 元空间的一个点，即一个 p 维向量，两个向量之间的距离记为 $d(x_i, x_j)$，满足如下条件：

1）非负性：任意两个向量间的距离非负，即

$d(x_i, x_j) \geqslant 0$，且 $d(x_i, x_j) = 0$ 当且仅当 $x_i = x_j$

2）对称性：　　$d(x_i, x_j) = d(x_j, x_i)$

3）三角不等式：　　$d(x_i, x_j) \leqslant d(x_i, x_k) + d(x_k, x_j)$

在聚类分析中最常用的是欧氏距离。

（1）欧氏距离

$$d(x_i, x_j) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2\right]^{1/2} \tag{6.1.1}$$

（2）绝对距离

$$d(x_i, x_j) = \sum_{k=1}^{p}|x_{ik} - x_{jk}| \tag{6.1.2}$$

（3）闵氏距离

$$d(x_i, x_j) = \left[\sum_{k=1}^{p}|x_{ik} - x_{jk}|^m\right]^{1/m} \tag{6.1.3}$$

其中 $m(m>0)$ 为常数。

（4）切氏距离

切比雪夫距离（Chebyshev distance）简称切氏距离，定义如下：

$$d(x_i, x_j) = \max_{1 \leqslant k \leqslant p}|x_{ik} - x_{jk}| \tag{6.1.4}$$

（5）方差加权距离

$$d(x_i, x_j) = \left[\sum_{k=1}^{p}(x_{ik} - x_{jk})^2 / s_k^2\right]^{1/2} \tag{6.1.5}$$

其中 $s_k^2 = \dfrac{1}{n-1}\sum_{j=1}^{n}(x_{jk} - \overline{x}_k)^2$，$\overline{x}_k = \dfrac{1}{n}\sum_{j=1}^{n}x_{jk}$。

（6）马氏距离

$$d(x_i, x_j) = (x_i - x_j)^{\mathrm{T}}\Sigma^{-1}(x_i - x_j) \tag{6.1.6}$$

其中 Σ 为样品的协方差矩阵。

在 MATLAB 中，计算距离的命令是 pdist，调用格式为：

```
Y= pdist(X,distance)
```

输入的 X 是一个矩阵，行为个体，列为指标，distance 是距离的类型。若省略 distance，则输出的 Y 是一个行向量，向量的长度为 $(N-1) * N/2$，其中 N 是样本的容量，Y 的元素分别为个体 $(1,2),(1,3),\cdots,(1,N),(2,3),\cdots(2,N),\cdots(N-1,N)$ 之间的欧氏距离。

可选项 distance 有：euclidean（欧氏距离）；cityblock（绝对距离）；minkowski（明氏距

离）（$m=2$）；chebychev（切氏距离）；seuclidean（方差加权距离）；mahalanobis（马氏距离）。

例 6.1.1 2008 年我国 5 省、区、市城镇居民人均年家庭收入见表 6-1，为了研究上述 5 个省、区、市的城镇居民收入差异，需要利用统计资料对其进行分类，指标变量有 4 个，计算各省、区、市之间的前 6 种距离。

表 6-1 5 省、区、市城镇居民人均家庭收入 （单位：元/人）

省（区、市）	工薪收入	经营净收入	财产性收入	转移性收入
北 京	18 738.96	778.36	452.75	7 707.87
上 海	21 791.11	1 399.14	369.12	6 199.77
安 徽	9 302.38	959.43	293.92	3 603.72
陕 西	9 794.82	544	151.46	3 356.85
新 疆	9 422.22	938.15	141.75	1 976.49

解：编写程序如下：

```
x=[18738.96      778.36      452.75      7707.87
   21791.11     1399.14      369.12      6199.77
    9302.38      959.43      293.92      3603.72
    9794.82      544         151.46      3356.85
    9422.22      938.15      141.75      1976.49];
d1=pdist(x);          % 计算各行之间的欧氏距离
```

为了得到距离矩阵，键入以下命令：

```
D=squareform(d1);     % 将行向量 d1 转变成一个方阵
```

输出结果为：

```
D=1.0e+004*
     0      0.3462      1.0293      0.9954      1.0944
0.3462           0      1.2763      1.2360      1.3080
1.0293      1.2763           0      0.0704      0.1639
0.9954      1.2360      0.0704           0      0.1483
1.0944      1.3080      0.1639      0.1483           0
```

矩阵 D 中第 i 行 j 列的元素表示 x 中的第 i 个个体与第 j 个个体之间的欧氏距离。如矩阵 D 中的第 3 行第 2 列为 12 763，表示上海与陕西的欧氏距离为 12 763，其余类推。

若想得到下三角阵，则有命令：

```
S=tril(squareform(d1))        % 提取方阵 squareform(d1)的下三角部分
```

输出结果为：

```
S=1.0e+004*
         0         0         0         0         0
    0.3462         0         0         0         0
    1.0293    1.2763         0         0         0
    0.9954    1.2360    0.0704         0         0
    1.0944    1.3080    0.1639    0.1483         0
```

```
d2=pdist(x,'cityblock');          % 计算绝对距离
D2=squareform(d2)
```

输出结果为：

```
D2=1.0e+004*
```

```
     0        0.5265        1.3881        1.3831        1.5519
0.5265             0        1.5600        1.5912        1.7281
1.3881        1.5600             0        0.1297        0.1921
1.3831        1.5912        0.1297             0        0.2157
1.5519        1.7281        0.1921        0.2157             0
d3= pdist(x,'minkowski',3);    % 计算明氏距离，d3为1行10列的行向量
d4= pdist(x,'chebychev')       % 计算切氏距离
d5= pdist(x,'seuclidean')      % 计算方差加权距离
d6= pdist(x,'mahalanobis')     % 计算马氏距离
```

欧氏距离与量纲有关，因此，有时需要对数据进行预处理，如标准化等，在MATLAB中的命令是zscore，调用格式为：

```
Z= zscore(X)
```

输入 X 表示 N 行 p 列的原始观测矩阵，行为个体，列为指标。输出 Z 为 X 的标准化矩阵，即 $Z=(X-\text{ones}(N,1)*\text{mean}(X))./(\text{ones}(N,1)*\text{std}(X))$，其中 mean($X$) 为行向量，表示各个指标的均值估计；std（$X$）表示指标的标准差估计；“./”表示对应元素相除；ones(N，1）表示元素全为1的行向量，向量的长度为 N。

聚类分析方法不仅可以对样品进行分类，而且可以对变量进行分类，在对变量进行分类时，常常采用相似系数来度量变量之间的相似性。

对 p 个指标变量进行聚类时，用相似系数来衡量变量之间的相似程度（关联度），若用 $C_{\alpha\beta}$ 表示变量 α、β 之间的相似系数，则应满足：

1）$|C_{\alpha\beta}|\leqslant 1$ 且 $C_{\alpha\alpha}=1$。

2）$C_{\alpha\beta}=\pm 1$ 当且仅当 $\alpha=k\beta$，$k\neq 0$。

3）$C_{\alpha\beta}=C_{\alpha\beta}$。

相似系数中最常用的是相关系数与夹角余弦。

例6.1.2 计算例6.1.1中各指标之间的相关系数与夹角余弦。

解：编写程序如下：

```
x= [⋯];                % 与例6.1.1数据相同
R= corrcoef(x);        % 指标之间的相关系数
```

输出结果为：

```
R=
   1.0000    0.6183    0.8138    0.8931
   0.6183    1.0000    0.4287    0.2927
   0.8138    0.4287    1.0000    0.9235
   0.8931    0.2927    0.9235    1.0000
x1= normc(x);          % 将x的各列化为单位向量
J= x1' * x1            % 计算夹角余弦
```

输出结果为：

```
J=
   1.0000    0.9536    0.9609    0.9797
   0.9536    1.0000    0.9026    0.8990
   0.9609    0.9026    1.0000    0.9833
   0.9797    0.8990    0.9833    1.0000
```

6.1.3 类间距离与递推公式

前面我们介绍了两个向量之间的距离，下面我们介绍两个类别之间的距离。设 d_{ij} 表示两个样品 x_i、x_j 之间的距离，G_p、G_q 分别表示两个类别，各自含有 n_p、n_q 个样品。

（1）最短距离

$$D_{pq} = \min_{i\in G_p, j\in G_q} d_{ij} \tag{6.1.7}$$

即用两类中样品之间的距离最短者作为两类间距离。

（2）最长距离

$$D_{pq} = \max_{i\in G_p, j\in G_q} d_{ij} \tag{6.1.8}$$

即用两类中样品之间的距离最长者作为两类间距离。

（3）类平均距离

$$D_{pq} = \frac{1}{n_p n_q}\sum_{i\in G_p}\sum_{j\in G_q} d_{ij} \tag{6.1.9}$$

即用两类中所有两两样品之间距离的平均作为两类间距离。

（4）重心距离

$$D_{pq} = d(\overline{x}_p, \overline{x}_q) = \sqrt{(\overline{x}_p - \overline{x}_q)^{\mathrm{T}}(\overline{x}_p - \overline{x}_q)} \tag{6.1.10}$$

其中 $\overline{x}_p$、$\overline{x}_q$ 分别是 G_p、G_q 的重心，这是用两类重心之间的欧氏距离作为类间距离。

（5）离差平方和距离（ward）

$$D^2_{pq} = \frac{n_p n_q}{n_p + n_q}(\overline{x}_p - \overline{x}_q)^{\mathrm{T}}(\overline{x}_p - \overline{x}_q) \tag{6.1.11}$$

显然，离差平方和距离与重心距离的平方成正比。

设有两类 G_p、G_q 合并成新的一类 G_r，包含了 $n_r = n_p + n_q$ 个样品，如何计算 G_r 与其他类别 $G_k(k \neq p,q)$ 之间的距离？这就需要建立类间距离的递推公式。

（1）最短距离

$$D_{rk} = \min(D_{pk}, D_{qk}) \tag{6.1.12}$$

（2）最长距离

$$D_{rk} = \max(D_{pk}, D_{qk}) \tag{6.1.13}$$

（3）类平均距离

$$D_{rk} = \frac{n_p}{n_r}D_{pk} + \frac{n_q}{n_r}D_{qk} \tag{6.1.14}$$

显然，G_k 与 G_r 之间类平均距离是 G_k 与 G_p 之间类平均距离以及 G_k 与 G_q 之间类平均距离的加权平均。

（4）重心距离

$$D^2_{rk} = \frac{n_p}{n_r}D^2_{pk} + \frac{n_q}{n_r}D^2_{qk} - \frac{n_p}{n_r}\frac{n_q}{n_r}D^2_{pq} \tag{6.1.15}$$

证明：

$$D^2_{rk} = \| \overline{x}_r - \overline{x}_k \|^2 = \left\| \frac{1}{n_p + n_q}\sum_{x_i\in(G_p+G_q)} x_i - \frac{1}{n_k}\sum_{x_i\in G_k} x_i \right\|$$

$$= \left\| \frac{1}{n_p + n_q}\sum_{x_i\in G_p} x_i + \frac{1}{n_p + n_q}\sum_{x_i\in G_q} x_i - \frac{1}{n_k}\sum_{x_i\in G_k} x_i \right\|^2$$

$$
\begin{aligned}
&= \left\| \frac{n_p}{n_r n_p} \sum_{x_i \in G_p} x_i + \frac{n_q}{n_r n_q} \sum_{x_i \in G_q} x_i - \frac{1}{n_k} \sum_{x_i \in G_k} x_i \right\|^2 \\
&= \left\| \frac{n_p}{n_r} \overline{x}_p + \frac{n_q}{n_r} \overline{x}_q - \overline{x}_k \right\|^2 \\
&= \left\| \frac{n_p}{n_r} \overline{x}_p + \frac{n_q}{n_r} \overline{x}_q - \frac{n_p}{n_r} \overline{x}_k - \frac{n_q}{n_r} \overline{x}_k \right\|^2 \\
&= \left\| \frac{n_p}{n_r} \overline{x}_p - \frac{n_p}{n_r} \overline{x}_k + \frac{n_q}{n_r} \overline{x}_q - \frac{n_q}{n_r} \overline{x}_k \right\|^2 \\
&= \frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2 - \frac{n_p n_q}{n_r^2} D_{pq}^2
\end{aligned}
$$

(5) 离差平方和距离

$$
D_{rk}^2 = \frac{n_p + n_k}{n_r + n_k} D_{pk}^2 + \frac{n_q + n_k}{n_r + n_k} D_{qk}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2 \tag{6.1.16}
$$

6.2 谱系聚类与K均值聚类

6.2.1 谱系聚类

谱系聚类法是目前应用较为广泛的一种聚类法。谱系聚类是根据生物分类学的思想对研究对象进行分类的方法。在生物分类学中，分类的单位是：门、纲、目、科、属、种，其中种是分类的基本单位，分类单位越小，它所包含的生物就越少，生物之间的共同特征就越多。利用这种思想，谱系聚类首先将每个样品看成一类，然后把最相似（距离最近或相似系数最大）的样品聚为小类，再将已聚合的小类按各类之间的相似性（用类间距离度量）进行再聚合，随着相似性的减弱，最后将一切子类都聚为一大类，从而得到一个按相似性大小聚结起来的谱系图。

1. 谱系聚类的步骤

谱系聚类的步骤如下：

1）n 个样品开始作为 n 个类，计算两两之间的距离或相似系数，得到实对称矩阵

$$
D_0 = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}
$$

2）从 D_0 的非主对角线上找最小（距离）或最大元素（相似系数），设该元素是 D_{pq}，则将 G_p、G_q 合并成一个新类 $G_r=(G_p,G_q)$，在 D_0 中去掉 G_p、G_q 所在的两行、两列，并加上新类 G_r 与其余各类之间的距离或相似系数，得到 $n-1$ 阶矩阵 D_1。

3）从 D_1 出发重复步骤 2）的做法得到 D_2，再由 D_2 出发重复上述步骤，直到全部样品聚为一个大类为止。

4）在合并过程中要记下合并样品的编号及两类合并时的水平，并绘制谱系聚类图。

例 6.2.1 从例 6.1.1 算得的样品间的欧氏距离矩阵出发，分别用最短距离与最长距离方法进行谱系聚类。

解： 我们用 1、2、3、4、5 分别表示北京、上海、安徽、陕西和新疆，将欧氏距离矩阵除以 10^4，记为 D_0：

$$D_0=\begin{array}{c} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{ccccc} \{1\} & \{2\} & \{3\} & \{4\} & \{5\} \end{array} \\ \begin{bmatrix} 0 & & & & \\ 0.346\,2 & 0 & & & \\ 1.029\,3 & 1.276\,3 & 0 & & \\ 1.157\,5 & 1.393\,2 & 0.142\,8 & 0 & \\ 1.094\,4 & 1.308\,0 & 0.163\,9 & 0.128\,0 & 0 \end{bmatrix} \end{array}$$

1）最短距离法：将各个样品看成一类，即 $G_i=\{i\}(i=1,2,3,4,5)$，从 D_0 可以看出各类中距离最短的是 $d_{54}=0.128\,0$，因此将 G_4、G_5 在 0.128 0 水平上合成一个新类 $G_6=\{4,5\}$，计算 G_6 与 G_1、G_2、G_3 之间的最短距离，得

$$D_{61}=\min\{d_{41},d_{51}\}=1.094\,4$$

$$D_{62}=\min\{d_{42},d_{52}\}=1.308\,0$$

$$D_{63}=\min\{d_{43},d_{53}\}=0.142\,8$$

将计算结果作为第一列，从 D_0 中去掉第 4、5 行与 4、5 列，剩余元素作为其余各列，得到 D_1：

$$D_1=\begin{array}{c} \\ \{4,5\} \\ \{1\} \\ \{2\} \\ \{3\} \end{array}\begin{array}{c} \begin{array}{cccc} \{4,5\} & \{1\} & \{2\} & \{3\} \end{array} \\ \begin{bmatrix} 0 & & & \\ 1.094\,4 & 0 & & \\ 1.308\,0 & 0.346\,2 & 0 & \\ 0.142\,8 & 1.029\,3 & 1.276\,3 & 0 \end{bmatrix} \end{array}$$

从 D_1 可以看到 G_6 与 G_3 的距离最小，因此在 0.142 8 的水平上将 G_6 与 G_3 合成一类 G_7，即 $G_7=\{3,4,5\}$，计算 G_7 与 G_1、G_2 之间的最短距离，得

$$D_{71}=\min\{D_{61},D_{51}\}=1.094\,4$$

$$D_{72}=\min\{D_{62},D_{52}\}=1.308\,0$$

将计算结果作为第 1 列，从 D_1 中划掉 {3，4} 与 {5} 所在的行与列，剩余元素作为其他列，得

$$D_2=\begin{array}{c} \\ \{3,4,5\} \\ \{1\} \\ \{2\} \end{array}\begin{array}{c} \begin{array}{ccc} \{3,4,5\} & \{1\} & \{2\} \end{array} \\ \begin{bmatrix} 0 & & \\ 1.094\,4 & 0 & \\ 1.308\,0 & 0.346\,2 & 0 \end{bmatrix} \end{array}$$

从 D_2 可以看出 G_1、G_2 最接近，在 0.346 2 的水平上合并成一类 G_8，至此只剩下两类 G_7、G_8，它们之间的距离为 1.029 3，故在此水平上将 G_7、G_8 合成一类，包含了全部的 5 个样品。

2）最长距离法：

$$D_0=\begin{array}{c} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{ccccc} \{1\} & \{2\} & \{3\} & \{4\} & \{5\} \end{array} \\ \begin{bmatrix} 0 & & & & \\ 0.346\,2 & 0 & & & \\ 1.029\,3 & 1.276\,3 & 0 & & \\ 1.157\,5 & 1.393\,2 & 0.142\,8 & 0 & \\ 1.094\,4 & 1.308\,0 & 0.163\,9 & 0.128\,0 & 0 \end{bmatrix} \end{array}$$

非对角线上最小元素为 0.128 0，在此水平上 $G_6=\{4,5\}$，计算 G_6 与 G_1、G_2、G_3 之间的最长距离，得

$$D_{61}=\max\{d_{41},d_{51}\}=1.157\,5$$

$$D_{62}=\max\{d_{42},d_{52}\}=1.3932$$

$$D_{63}=\max\{d_{43},d_{53}\}=0.1639$$

将计算结果作为第1列，从 D_0 中去掉第4、5行与4、5列，剩余元素作为其余各列，得到 D_1，

$$D_1=\begin{array}{c} \\ \{4,5\} \\ \{1\} \\ \{2\} \\ \{3\} \end{array}\begin{array}{c} \begin{array}{cccc} \{4,5\} & \{1\} & \{2\} & \{3\} \end{array} \\ \begin{bmatrix} 0 & & & \\ 1.1575 & 0 & & \\ 1.3932 & 0.3462 & 0 & \\ 0.1639 & 1.0293 & 1.2763 & 0 \end{bmatrix} \end{array}$$

从 D_1 可以看出 G_6 与 G_3 的距离最小，因此在0.1639的水平上将 G_6 与 G_3 合成一类 G_7，即 $G_7=\{3,4,5\}$，计算 G_7 与 G_1、G_2 之间的最长距离，得

$$D_{71}=\max\{D_{61},D_{51}\}=1.1575$$

$$D_{72}=\max\{D_{62},D_{52}\}=1.3932$$

将计算结果作为第1列，从 D_1 中划掉 {3，4} 与 {5} 所在的行与列，剩余元素作为其他列，得

$$D_2=\begin{array}{c} \\ \{3,4,5\} \\ \{1\} \\ \{2\} \end{array}\begin{array}{c} \begin{array}{ccc} \{3,4,5\} & \{1\} & \{2\} \end{array} \\ \begin{bmatrix} 0 & & \\ 1.1575 & 0 & \\ 1.3932 & 0.3462 & 0 \end{bmatrix} \end{array}$$

从 D_2 可以看出 G_1、G_2 最接近，在0.3462的水平上合并成一类 G_8，至此只剩下两类 G_7、G_8，它们之间的最长距离为1.3932，故在此水平上将 G_7、G_8 合成一类，包含了全部的5个样品。

2. 谱系聚类的 MATLAB 实现

为了方便快捷地实现大样本的聚类分析，我们利用 MATLAB 软件实现谱系聚类。

1）谱系聚类命令为 linkage，其调用格式为：

```
Z= linkage(Y,method)
```

输入 Y 是一个距离矩阵，例如 Y 是由 pdist 命令生成的欧氏距离向量。

method 是一个可选项，如最长距离、最短距离等。

'single'——最短距离（默认状态）

'complete'——最长距离

'average'——类平均距离

'weighted'——加权平均距离

'centroid'——重心距离

'ward'——离差平方和距离

输出 Z 是一个（$N-1$）行3列矩阵，Z 的第1列和第2列均为正整数，第3列表示聚类的水平，每一行表示在相同的聚类水平上将个体合并成新的一类，每生成一个新的类，其编号将在现有基础上增加1。

2）作谱系聚类图命令为 dendrogram，其调用格式为：

```
H= dendrogram(z,N)
```

输入 Z 是一个（$N-1$）行3列的矩阵，由 linkage 命令生成，N 是样本容量。输出产生一个树谱系聚类图，每两类通过线段连接，高度表示类间的距离。此命令作出 m 个样本的图形，

缺省时默认为 30。

3）输出聚类结果命令为 cluster，其调用格式为：

```
T= cluster(z,k)
```

输入 Z 是一个（$N-1$）行 3 列的矩阵，由 linkage 命令生成，N 是样本容量；k 是分类数目；输出 T 是一个列向量（N 行 1 列），每一个元素均为正整数，且最大的数字不超过 k，第 i 行的数字 1 表示第 i 个个体属于第 1 类。

如果遇到大样本数据，为了便于得到每一类样本的编号，可以利用如下命令：

```
find(T== 1)          % 找出属于第 1 类的样品编号
```

例 6.2.2 利用 MATLAB 软件对例 6.1.1 中的 5 个省、区、市进行聚类分析。

解： 编写程序如下：

```
% 输入样本数据
x= [18738.96    778.36    452.75    7707.87
    21791.11   1399.14    369.12    6199.77
     9302.38    959.43    293.92    3603.72
     8354.63    638.76     65.33    2610.61
     9422.22    938.15    141.75    1976.49];
d= pdist(x);              % 欧氏距离
z1= linkage(d)            % 类间距离为最短距离
```

输出结果为：

```
z1=
1.0e+ 004 *
      0.0004    0.0005    0.1280   % 在 0.1280 的水平，G4，G5 合成一类为 G6
      0.0003    0.0006    0.1428   % 在 0.1428 的水平，G6，G3 合成一类为 G7
      0.0001    0.0002    0.3462   % 在 0.3462 的水平，G1，G2 合成一类为 G8
      0.0007    0.0008    1.0293   % 在 10.0293 的水平，G7，G8 合成一类
H= dendrogram(z1)      % 作谱系聚类图
```

输出图形如图 6-1 所示。

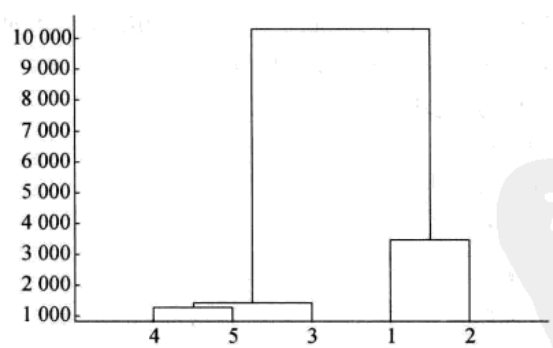


图 6-1 最短距离聚类图

```
z2= linkage(d,'complete')      % 选择类间距离为最长距离时
```

输出结果为：

```
z2=
1.0e+ 004 *
      0.0004    0.0005    0.1280   % 在 0.1280 的水平，G4，G5 合成一类为 G6
      0.0003    0.0006    0.1639   % 在 0.1639 的水平，G6，G3 合成一类为 G7
```

```
    0.0001    0.0002    0.3462  % 在 0.3462 的水平,G1,G2 合成一类为 G8
    0.0007    0.0008    1.3932  % 在 10.3932 的水平,G7,G8 合成一类
H2= dendrogram(z2)    % 作谱系聚类图
```

输出图形如图 6-2 所示。

```
z3= linkage(d,'average')    % 选择类间距离为类平均距离时
```

输出结果为：

```
z3=
1.0e+ 004 *
    0.0004    0.0005    0.1280  % 在 0.1280 的水平,G4,G5 合成一类为 G6
    0.0003    0.0006    0.1533  % 在 0.1533 的水平,G6,G3 合成一类为 G7
    0.0001    0.0002    0.3462  % 在 0.3462 的水平,G1,G2 合成一类为 G8
    0.0007    0.0008    1.2098  % 在 10.2098 的水平,G7,G8 合成一类
H= dendrogram(z3)      % 作谱系聚类图
```

输出图形如图 6-3 所示。

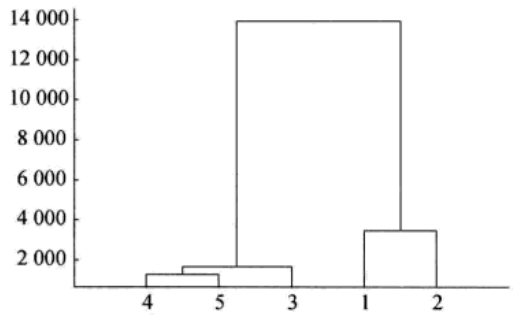


图 6-2　最长距离聚类图

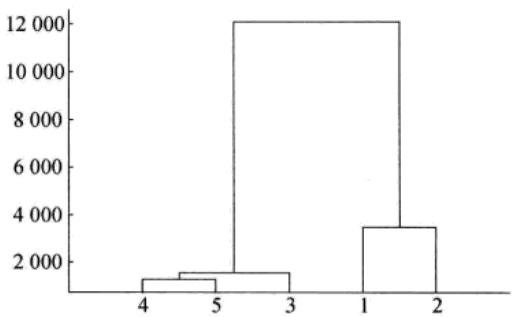


图 6-3　类平均距离聚类图

若我们不知道实际的观测数据，但已经知道样品之间的距离，那么如何在 MATLAB 中实现相应的聚类？

例 6.2.3　欧州各国语言有许多相似之处，有的十分相近，以 E、N、Da、Du、G、Fr、S、I、P、H、Fi 分别表示英语、挪威语、丹麦语、荷兰语、德语、法语、西班牙语、意大利语、波兰语、匈牙利语和荷兰语 11 种语言。人们以任两种语言对 1～10 这 10 个数字拼写中第一个字母不相同的个数定义两种语言间的“距离”。这种距离是广义距离，例如，英语和挪威语只有数字 1 和 8 的第一个字母不同，故这两种语言间的距离定义为 2。这样得到 11 种语言间的距离矩阵如下：

	E	N	Da	Du	G	Fr	S	I	P	H	Fi
E	0										
N	2	0									
Da	2	1	0								
Du	7	5	6	0							
G	6	4	5	5	0						
Fr	6	6	6	9	7	0					
S	6	6	5	9	7	2	0				
I	6	6	5	9	7	1	1	0			
P	7	7	6	10	8	5	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
Fi	9	9	9	9	9	9	9	9	9	8	0

解：编写程序如下：

```
d= [2 2 7 6 6 6 7 9 9 1 5 4 6 6 6 7 8 9 6 5 6 5 5 5 6 8 9 5 9 9 9 10 8 9 7 7 7 8 9 9 2 1 5 10 9 1 3 10 9 4 10 9 10
    9 8];                          % 按列输入距离矩阵(只输入下三角阵中的非零元素)
z4= linkage(d,'centroid');         % 重心距离
H2= dendrogram(z4)                 % 谱系图
z5= linkage(d,'ward');             % 离差平方和距离
figure(2)
H3= dendrogram(z5)          % 谱系图
```

重心距离的谱系聚类图如图 6-4 所示。离差平方和距离的谱系聚类图如图 6-5 所示。

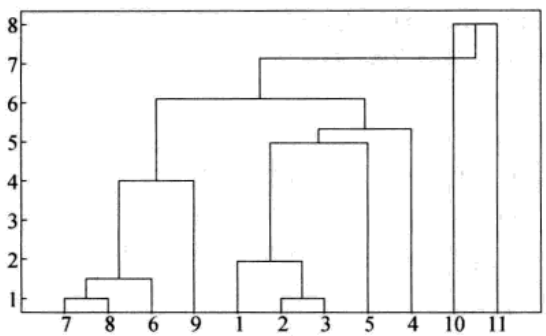


图 6-4　重心距离的谱系聚类图

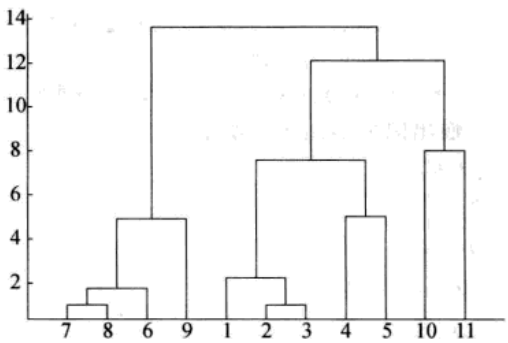


图 6-5　离差平方和距离的谱系聚类图

例 6.2.4　R. A. Fisher 在 1936 年发表的 Iris 数据中，研究某植物的萼片长、宽及花瓣长、宽。$x1$：萼片长，$x2$：萼片宽，$x3$：花瓣长，$x4$：花瓣宽。Iris 数据保存在 MATLAB 软件系统的文件 fisheriris. mat 中，用 meas 存储了取自三个总类 $G1$、$G2$ 和 $G3$，每一类取 50 个样本。试利用谱系聚类对 Iris 数据进行聚类。

解：从 MATLAB 系统中导入样本数据的命令为 load fisheriris。程序如下：

```
load   fisheriris          % 导入萼片的相关数据
d= pdist(meas)             % 计算欧氏距离
z1= linkage(d)             % 类间为最短距离
T= cluster(z1,3)           % 分为 3 类
g1= find(T== 1)            % 第一类里的样品编号
g2= find(T== 2)            % 第二类里的样品编号
g3= find(T== 3)            % 第三类里的样品编号
```

作原数据的两两指标的三个总类散点图如图 6-6 所示。程序如下：

```
subplot(2,3,1)
plot (meas(1:50,1),meas(1:50,2),'*',meas(51:100,1),meas(51:100,2),'g*',meas(101:150,1),meas
     (101:150,2),'ro')
title('x1- x2')
subplot(2,3,2)
plot (meas(1:50,1),meas(1:50,3),'*',meas(51:100,1),meas(51:100,3),'g*',meas(101:150,1),meas
     (101:150,3),'ro')
title('x1- x3')
subplot(2,3,3)
plot (meas(1:50,1),meas(1:50,4),'*',meas(51:100,1),meas(51:100,4),'g*',meas(101:150,1),meas
     (101:150,4),'ro')
```

```
title('x1- x4')
subplot(2,3,4)
plot (meas(1:50,2),meas(1:50,3),'*',meas(51:100,2),meas(51:100,3),'g*',meas(101:150,2),meas
     (101:150,3),'ro')
title('x2- x3')
subplot(2,3,5)
plot (meas(1:50,2),meas(1:50,4),'*',meas(51:100,2),meas(51:100,4),'g*',meas(101:150,2),meas
     (101:150,4),'ro')
title('x2- x4')
subplot(2,3,6)
plot (meas(1:50,3),meas(1:50,4),'*',meas(51:100,3),meas(51:100,4),'g*',meas(101:150,3),meas
     (101:150,4),'ro')
title('x3- x4')
```

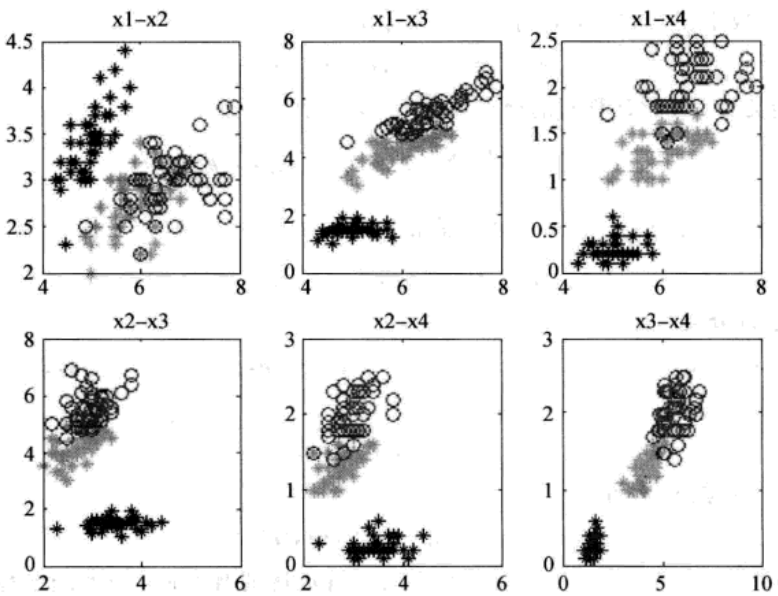


图 6-6 Iris 数据的散点图（两两指标）

结果显示第一类里只有两个样品，说明聚类效果不理想，为了提高聚类效果的优良性，我们将运用 K 均值聚类和模糊 C 均值聚类对其进行讨论。

同理可作两两指标经聚类分类后的数据散点图，从图形也可知聚类效果不理想。

6.2.2 K 均值聚类

谱系聚类法是先将每个样品看成一类，通过比较距离的大小逐步扩充类，因此，对于给定的数据，谱系聚类一定能够将样品合并为一类，分类的结果唯一，但是谱系聚类有一个缺点，样品一旦被分到某一类中就不能改变，且当样本容量较大时，计算量也相应地变大。克服此缺点的一个方法就是 K 均值聚类法，又称快速聚类法或动态聚类法。

在运用 K 均值聚类法之前，要根据实际问题先确定分类数 k，在每一类中选择有代表性的样品，这样的样品称为聚点。选择聚点的方法通常有最小最大原则。

若将 n 个样品分成 k 类，则先选择所有样品中距离最远的两个样品 x_{i1}、x_{i2} 为聚点，

使得

$$d(x_{i1},x_{i2}) = d_{i1i2} = \max\{d_{ij}\}$$

然后选择第 3 个聚点 x_{i3}，使得 x_{i3} 与前两个聚点的距离最小者等于所有其余的与 x_{i1}、x_{i2} 的较小距离中最大的，即

$$\min\{d(x_{i3},x_{ir}),r=1,2\} = \max\{\min\{d(x_j,x_{ir}),r=1,2\},j\neq i1,i2\}$$

最后按相同的原则选取 x_{ik}，重复前面的步骤，直至确定 k 个聚点 x_{i1}，x_{i2}，…，x_{ik}。

K 均值聚类的步骤（样品之间的距离采用欧氏距离）如下：

1）设第 k 个初始聚点的集合是

$$L^{(0)} = \{x_1^{(0)},x_2^{(0)},\cdots,x_k^{(0)}\}$$

记

$$G_i^{(0)} = \{x:d(x,x_i^{(0)}) \leqslant d(x,x_j^{(0)}),j=1,2,\cdots,k,j\neq i\} \quad (i=1,2,\cdots,k)$$

于是，将样品分成不相交的 k 类，得到一个初始分类

$$G^{(0)} = \{G_1^{(0)},G_2^{(0)},\cdots,G_k^{(0)}\}$$

2）从初始类 $G^{(0)}$ 开始计算新的聚点集合 $L^{(1)}$，计算

$$x_i^{(1)} = \frac{1}{n_i}\sum_{x_l\in G_i^{(0)}} x_l \quad (i=1,2\cdots,k)$$

其中 n_i 是类 $G^{(0)}$ 中的样品数，得到一个新的集合

$$L^{(1)} = \{x_1^{(1)},x_2^{(1)},\cdots,x_k^{(1)}\}$$

从 $L^{(1)}$ 开始再进行分类，记

$$G_i^{(1)} = \{x:d(x,x_i^{(1)}) \leqslant d(x,x_j^{(1)}),j=1,2,\cdots,k,j\neq i\} \quad (i=1,2,\cdots,k)$$

得到一个新的类

$$G^{(1)} = \{G_1^{(1)},G_2^{(1)},\cdots,G_k^{(1)}\}$$

3）重复上述步骤 m 次得

$$G^{(m)} = \{G_1^{(m)},G_2^{(m)},\cdots,G_k^{(m)}\}$$

其中 $x_i^{(m)}$ 是类 $G_i^{(m-1)}$ 的重心。$x_i^{(m)}$ 不一定是样品。当 m 逐渐增大时，分类趋于稳定。同时 $x_i^{(m)}$ 可以近似地看作 $G_i^{(m)}$ 的重心，即 $x_i^{(m+1)}\approx x_i^{(m)}$，$G_i^{(m+1)}\approx G_i^{(m)}$，此时结束计算。实际计算时，若对某一个 m，

$$G^{(m+1)} = \{G_1^{(m+1)},G_2^{(m+1)},\cdots,G_k^{(m+1)}\}$$

与

$$G^{(m)} = \{G_1^{(m)},G_2^{(m)},\cdots,G_k^{(m)}\}$$

相同，则结束计算。

MATLAB 软件中实现 K 均值聚类的命令是 kmeans，其调用格式为：

```
IDX= kmeans(X,K)
```

其功能是将原始数据矩阵 X 聚成 K 类，使得样本到类重心距离和最小，其中使用欧氏平方距离。输入 X 为原始观测数据，行为个体，列为指标。输出 IDX 为 N 行 1 列的列向量，包含每个样品属于哪一类的信息，类似于 cluster 的输出结果。

例 6.2.5 从 12 个不同地区测得了某树种的平均发芽率 x_1 与发芽势 x_2，数据见表 6-2，采用欧氏距离，将这 12 个地区以树种发芽情况按 K 均值聚类法聚为 2 类。

表 6-2 12 个地区某树种发芽情况

地　区	1	2	3	4	5	6	7	8	9	10	11	12
x_1	0.707	0.600	0.693	0.717	0.688	0.533	0.877	0.513	0.815	0.633	0.740	0.777
x_2	0.385	0.433	0.505	0.343	0.605	0.380	0.713	0.353	0.675	0.465	0.580	0.723

解：利用 MATLAB 软件中的命令 kmeans，可以实现 K 均值聚类。程序如下：

```
y=[.707 .6 .693 .717 .688 .533 .877 .513 .815 .633 .74 .777;
   .385 .433 .505 .343 .605 .38 .713 .353 .675 .465 .58 .723];
x=y';                          % 矩阵 x 的行为个体,列为指标
[a,b]=kmeans(x,2)              % 分为 2 类,输出 a 为聚类的结果,b 为聚类重心,每一行表示一个类的重心
```

输出结果：

```
a=
   2
   2
   2
   2
   1
   2
   1
   2
   1
   2
   1
   1
b=
   0.7794   0.6592
   0.6280   0.4091
% 提取样品
x1=x(find(a==1),:)     % 提取第 1 类里的样品
x2=x(find(a==2),:)     % 提取第 2 类里的样品
x1=
    0.6880   0.6050
    0.8770   0.7130
    0.8150   0.6750
    0.7400   0.5800
    0.7770   0.7230
x2=
    0.7070   0.3850
    0.6000   0.4330
    0.6930   0.5050
    0.7170   0.3430
    0.5330   0.3800
    0.5130   0.3530
    0.6330   0.4650
sd1=std(x1)
```

```
sd2= std(x2)                    % 分别计算第 1 类和第 2 类的标准差
sd1=
    0.0719    0.0641
sd2=
    0.0831    0.0603
plot(x(a== 1,1),x(a== 1,2),'r.',x(a== 2,1),x(a== 2,2),'b.')    % 作出聚类的散点图
```

分类结果的散点图，如图 6-7 所示。

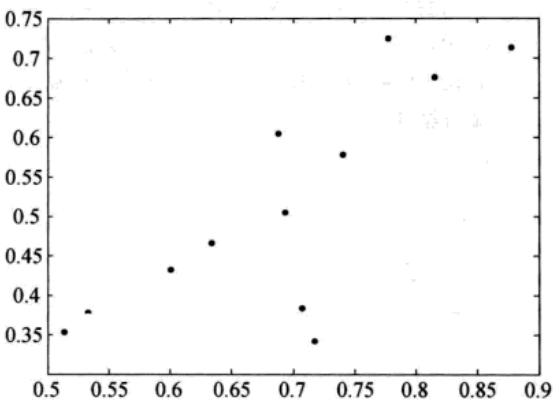


图 6-7　分类结果的散点图

例 6.2.6（续例 6.1.1）　利用 K 均值聚类对 5 个省、区、市进行聚类分析。

解：编写程序如下：

```
x= [⋯]                          % 输入数据，行为个体，列为指标
[a,b]= kmeans(x,3)    % 分为 3 类
```

输出结果：

```
a=
   1
   1
   2
   3
   3
b= 1.0e+ 004 *
    2.0265    0.1089    0.0411    0.6954
    0.9302    0.0959    0.0294    0.3604
    0.8888    0.0788    0.0104    0.2294
```

说明北京和上海为一类，安徽为一类，陕西和新疆为一类。

例 6.2.7（续例 6.2.4）　利用 K 均值聚类法将 Fisher 的 Iris 数据分为 3 类。

解：编写程序如下：

```
load fisheriris
[a b]= kmeans(meas,3)
```

输出结果：

```
a=
   1    1    1    1    1    1    1    1    1    1⋯
   1    1    1    1    1    1    1    1    1    1⋯
   1    1    1    1    1    1    1    1    1    1⋯
   1    1    1    1    1    1    1    1    1    1⋯
   1    1    1    1    1    1    1    1    1    1⋯
   3    3    2    3    3    3    3    3    3    3⋯
   3    3    3    3    3    3    3    3    3    3⋯
   3    3    3    3    3    3    3    2    3    3⋯
   3    3    3    3    3    3    3    3    3    3⋯
   3    3    3    3    3    3    3    3    3    3⋯
   2    3    2    2    2    2    3    2    2    2⋯
   2    2    2    3    3    2    2    2    2    3⋯
   2    3    2    3    2    2    3    3    2    2⋯
```

```
    2    2    2    3    2    2    2    2    3    2…
    2    2    3    2    2    2    3    2    2    3
b=
    5.0060    3.4280    1.4620    0.2460
    6.8500    3.0737    5.7421    2.0711
    5.9016    2.7484    4.3935    1.4339
n1= length(find(a== 1))   % 第1类的样品数
n2= length(find(a== 2))   % 第2类的样品数
n3= length(find(a== 3))   % 第3类的样品数
n1= 50,n2= 38,n3= 62
```

由此可见，K 均值聚类的效果比谱系聚类的效果好，但与实际的分类情况相比，K 均值聚类的结果依然不甚理想。

6.3 模糊均值聚类

本节我们将简述两种常用的模糊聚类方法：模糊 C 均值聚类和模糊减法聚类。

6.3.1 模糊 C 均值聚类

模糊 C 均值聚类（fuzzy c-mean cluster，FCM）是硬 C 均值聚类的推广，硬划分是指一个样品要么属于指定的类，要么不属于该类，二者必居其一。而模糊聚类则放松此要求，即样品以一定的概率属于某个指定类。

设 $X=\{x_1,x_2,\cdots,x_n\}\subset R^s$ 为样品集，n 为样本容量。将 X 分成 c 类，等价于将集合 X 表示成

$$X=X_1\cup X_2\cup\cdots\cup X_c \text{ 且 } X_i\cap X_j=\varphi, i\neq j$$

设 u_{ij} 是第 j 个样品属于第 i 个中心的隶属度，则

$$u_{ij}=\begin{cases}1 & x_j\in X_i\\ 0 & x_j\notin X_i\end{cases}$$

其中 $j=1,2,\cdots,n,i=1,2,\cdots,c$。

$U=(u_{ij})$ 是一个 $c\times n$ 的矩阵，称为隶属度矩阵或特征矩阵，其中每一列的元素只有一个 1，其余全部为 0。

硬划分的一个延拓是将隶属度矩阵定义为：

$$\sum_{i=1}^{c}u_{ij}=1(u_{ij}\geqslant 0)$$

此时的聚类方法称为模糊 C 均值聚类。

模糊 C 均值聚类通过求解如下的优化问题：

$$\text{Minimize}\quad J_m(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^m\parallel x_i-v_j\parallel^2 \tag{6.3.1}$$

其中 $v=\{v_1,v_2,\cdots,v_c\}\subset R^s(1<c<n)$ 是聚类中心；$m>1$ 是加权指数；m 的取值能够影响聚类的效果。

通过求解如下的方程得到聚类中心和隶属度矩阵：

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij})^m x_j}{\sum_{j=1}^{n}(u_{ij})^m} \quad (1 \leqslant i \leqslant c) \tag{6.3.2}$$

$$u_{ij} = \left[\sum_{k=1}^{c}\left(\frac{\| x_j - v_i \|^2}{\| x_j - v_k \|^2}\right)^{1/(m-1)}\right]^{-1} \quad (1 \leqslant i \leqslant c, 1 \leqslant j \leqslant n) \tag{6.3.3}$$

上述求解过程是一个不断重复的过程，直到达到控制误差范围之内。

具体的求解步骤如下：

1）预先给定分类数 c（如何选择合适的分类数将在聚类的有效性中详细讨论）和加权指数 m，初始化隶属度矩阵 $U=(u_{ij})$ 使得 $\sum_{i=1}^{c} u_{ij} = 1$。

2）用公式（6.3.2）计算聚类中心 $v_i, i=1,2,\cdots,c$。

3）根据公式（6.3.3）计算新的隶属度矩阵。

4）若 $J_m(U, V)$ 小于预先给定的正数 ε，则聚类过程结束，否则，转到步骤 2）。

MATLAB 软件里实现模糊 C 均值聚类的命令是 fcm，其调用格式为：

```
[center,U,obj_fcn]= fcm(data,n_cluster)
```

输入 data 为原始观测数据，行为个体，列为指标，n_cluster 为预先给定的聚类数。输出 center 是一个 n_cluster 行 p 列的矩阵，每 i 行表示第 i 类的重心。

U 是隶属度矩阵（n_cluster 行 N 列），每列的元素之和均为1，$U(i,j)$ 表示第 j 个个体属于第 i 列的隶属度。obj_fcn 是一个列向量，在每次计算过程中均使用公式（6.3.1）。

例 6.3.1 用模糊 C 均值聚类法对 Fisher 的 Iris 数据进行分类。

解： 编写程序如下：

```
load fisheriris                    % 导入 Iris 数据
[center u]= fcm(meas,3);           % meas 为 150 行 4 列的 3 个总体的观测数据
index1= find(u(1,:)== max(u))      % 寻找属于第 1 类的样品
index2= find(u(2,:)== max(u))      % 寻找属于第 2 类的样品
index3= find(u(3,:)== max(u))      % 寻找属于第 3 类的样品
index1=
52  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76
77  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99  100
102  107  114  120  122  124  127  128  134  139  143  147  150
index2=
51  53  78  101  103  104  105  106  108  109  110  111  112  113  115  116  117  118  119
121  123  125  126  129  130  131  132  133  135  136  137  138  140  141  142  144  145  146
148  149
index3=
1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24  25  26
27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50
```

从聚类的结果来看，只有第 3 类与预先给定的完全一致，其余两类均与实际的分类情况相差较大，因此误判率较高（参考本章 6.4 节），误判率为 16/150=0.106 7。

上述例子使用的距离均为欧氏距离，加权指数 $m=2$，对于例 6.3.1，若选择 $m=3$，则误判率为 15/150=0.1。只需要将原程序中的命令[center u]=fcm(x,3)修改为[center u]=fcm(x,3,3)即可。

6.3.2 模糊减法聚类

模糊C均值聚类的前提条件是需要知道分类数 c，如果对于分类数没有什么先验信息，那么我们就可以运用模糊减法聚类以确定相应的分类数和聚类中心，相应地该聚类数及聚类中心可以应用到模糊C均值聚类，因此，模糊减法聚类可以看作是模糊C均值聚类的前期工作。

设 $X=\{x_1,x_2,\cdots,x_n\}\subset R^s$ 为样品集，n 为样本容量。模糊减法聚类认为每个样品均为潜在的聚类中心，令 $d(v_i,x_j)^2=\|v_i-x_j\|^2$ 表示聚类中心 v_i 与样品 x_j 之间欧氏距离平方，在 v_i 处的爬山函数（mountain function）定义为

$$M(v_i)=\sum_{j=1}^{n}e^{-\alpha\|v_i-x_j\|^2}$$

其中 α 是一个正常数。爬山函数的取值越大，说明聚类中心 v_i 与样品的距离越小，因此，我们选择那些能够使得爬山函数取得较大值的 v_i 作为聚类中心。

令 M_1^* 是爬山函数的最大值，即 $M_1^*=\max(M(v_i))$，同时令 M_1^* 对应的中心为 v_1^*，于是，v_1^* 为第一个聚类中心，为了寻找其他的聚类中心，有必要消除 v_1^* 对聚类的影响，因此，考虑如下的函数

$$\hat{M}^j(v_i)=\hat{M}^{j-1}(v_i)-\hat{M}_{j-1}^*(v_i)\sum_{j=1}^{n}e^{-\beta\|v_i-v_{j-1}^*\|^2}$$

其中，$\hat{M}^j(v_i)$为新的爬山函数，$\hat{M}^{j-1}(v_i)$为上一步的爬山函数，$\hat{M}_{j-1}^*(v_i)$是$\hat{M}^{j-1}(v_i)$的最大值，v_{j-1}^*是新的聚类中心，β是一个正常数。

由于爬山法的计算量较大，Chiu对上述的爬山函数进行改进，定义一个新的爬山函数，

$$M(x_i)=\sum_{j=1}^{n}e^{-\alpha\|x_i-x_j\|^2}$$

其中 α 是一个正常数。

令 M_1^* 是爬山函数的最大值，即 $M_1^*=\max(M(x_i))$，同时令 M_1^* 对应的点为 x_1^*，于是，x_1^* 为第一个聚类中心，为了寻找其他的聚类中心，有必要消除 x_1^* 对聚类的影响，因此，考虑如下的函数

$$\hat{M}^j(x_i)=\hat{M}^{j-1}(x_i)-\hat{M}_{j-1}^*(x_i)\sum_{j=1}^{n}e^{-\beta\|x_i-x_{j-1}^*\|^2}$$

其中，$\hat{M}^j(x_i)$为新的爬山函数，$\hat{M}^{j-1}(x_i)$为上一步的爬山函数，$\hat{M}_{j-1}^*(x_i)$是$\hat{M}^{j-1}(x_i)$的最大值，x_{j-1}^*是新的聚类中心，β是一个正常数。

MATLAB中模糊减法聚类的命令是subclust，调用格式为：

```
C= subclust(X,RADII)
```

输入 X 为原始观测数据，行为个体，列为指标；RADII为介于0、1之间的数，通常为0.2～0.5，值越小，聚类中心的容量就越大。输出C运用模糊减法聚类方法的聚类中心的估计。

例6.3.2 用模糊减法聚类法确定Iris数据的聚类中心。

解： 我们取RADII=0.6编写程序如下：

```
load fisheriris          % 导入 Iris 数据
c= subclust(meas,0.6)
c=
   6.0000    2.9000    4.5000    1.5000
   5.0000    3.4000    1.5000    0.2000
   6.8000    3.0000    5.5000    2.1000
```

6.4 聚类的有效性

6.4.1 谱系聚类的有效性

在研究最佳聚类数之前，我们先讨论样品之间和两个总体（类）之间究竟采用何种距离为好。先假定样品之间的距离已定，例如选取欧氏距离。对于类间的 5 种不同距离，哪一种距离使得聚类的效果最好呢？为此我们计算 cophenet 相关系数。聚类树的 cophenet 距离与生成该聚类树的原始距离之间的线性相关系数定义为聚类树的 cophenet 相关，因此，它度量了个体间的不相似性，该系数越接近于 1，则聚类效果越好。

在 MATLAB 中计算 cophenet 相关系数的命令为：

```
R= cophenet(z,d)
```

其中，z 是用某种类间距离 linkage 后的结果，d 是样品之间的某种距离。

例 6.4.1（续例 6.1.1） 2008 年我国 5 省、区、市城镇居民人均年家庭收入见表 6-1，在进行谱系聚类时，选择哪种类间距离最好？

解：以样品间的距离为欧氏距离为例，考虑类间的 5 种不同距离：

最短距离：`z1= linkage(d)`

最长距离：`z2= linkage(d,'complete')`

类平均距离：`z3= linkage(d,'average')`

重心距离：`z4= linkage(d,'centroid')`

离差平方和：`z5= linkage(d,'ward')`

其中 d=pdist(x)，x 为原始矩阵。

```
R= [cophenet(z1,d),cophenet(z2,d),cophenet(z3,d),cophenet(z4,d),cophenet(z5,d)]  % 计算 cophenet 相关系数
```

输出结果如下：

```
R=
   0.9809    0.9811    0.9812    0.9812    0.9803
```

由于最大值为 0.981 2，所以类间距离为类平均距离和重心距离时效果最好。如果我们要找到最理想的分类方法，可以对每一种样品之间的距离都计算上述的复合相关系数，这样就可以找到最理想的样品距离与对应的类间距离。

以上我们考虑了样品之间距离与类间距离如何搭配可以使得聚类效果最好，但究竟分为几类最合适，到目前为止，还没有从理论上完全解决这一问题，通常的准则有：

（1）R^2 统计量

$$R_k^2 = \frac{B_k}{T} = 1 - \frac{P_k}{T}$$

假定已将 n 个样品分为 k 类，记为 G_1，G_2，…，G_k；n_t 表示 G_t 类的样品个数（$n_1+n_2+\cdots+n_k=n$）；$\overline{x}^t$ 表示 G_t 的重心，即 $\overline{x}^t=\frac{1}{n_t}(x_1^t+\cdots+x_{n_t}^t)$，则 G_t 类中 n_t 个样品的离差平方和为 $W_t=\sum_{i=1}^{n_t}(x_i^t-\overline{x}^t)^{\mathrm{T}}(x_i^t-\overline{x}^t)$。

所有样品的总离差平方和为：

$$T=\sum_{t=1}^{k}\sum_{i=1}^{n_t}(x_i^t-\overline{x})^{\mathrm{T}}(x_i^t-\overline{x})$$

$$\bar{x}=\frac{1}{n}(x_1+x_2+\cdots+x_n)$$

T 可以分解为：

$$T=\sum_{t=1}^{k}\sum_{i=1}^{n_t}(x_i^t-\bar{x}^t+\bar{x}^t-\bar{x})^{\mathrm{T}}(x_i^t-\bar{x}^t+\bar{x}^t-\bar{x})$$

$$=\sum_{t=1}^{k}W_t+\sum_{t=1}^{k}\sum_{i=1}^{n_t}n_t(\bar{x}^t-\bar{x})^{\mathrm{T}}(\bar{x}^t-\bar{x})=P_k+B_k$$

令 $R_k^2=\frac{B_k}{T}=1-\frac{P_k}{T}$，则 R_k^2 值越大，也就是 B_k/T 越大，表示 k 个类的类间偏差平方和的总和 B_k 在总离差平方和 T 中所占的比例越大，说明 k 个类能够区分开。因此统计量 R_k^2 可用于评价合并为 k 个类时的聚类效果，R_k^2 越大，聚类效果越好。

当样品各自为一类时 $R^2=1$，而当所有的样品为同一类时，$R^2=0$，因此如何恰当地使用该准则，要具体问题具体分析。由于 R^2 随着 n 的减少而减少，可以从 R^2 取值的变化来确定分为几类比较合适。

例 6.4.2 试利用 R^2 统计量确定 Iris 数据的分类数。

解： 编写程序如下：

```
load fisheriris
x= meas;[n,p]= size(x);n1= n- 1;
format long
c= zeros(n1- 1,1);
for j= 2:n1
      d= pdist(x);
      z1= linkage(d,'complete');
      c= cluster(z1,j);
      k= 1;
      if k<= j
            b= find(c== k);  l= length(b)- 1;
            if b> 0
            a= x(b,:);
            c(j)= sum(l * var(a))+ c(j);
            end
      end
end
R= 1- c/sum(n1 * var(x));
optimaln= find(R== max(R))
```

输出结果为：

```
optimaln=
          8
```

(2) 伪 F 统计量

$$F=\frac{(T-P_k)/(k-1)}{P_k/(n-k)}=\frac{B_k}{P_k}\frac{n-k}{k-1}$$

伪 F 统计量用于评价分为 k 类的效果。伪 F 统计量的值越大表示这 n 个样品越可显著地分为 k 类。

(3) 伪 t^2 统计量

$$t^2=\frac{B_{KL}^2}{(W_K+W_L)/(n_K+n_L-2)}$$

其中 W_K、W_L 分别表示第 K 类和第 L 类的离差平方和，$B_{KL}^2=W_M-(W_K+W_L)$ 表示合并类 G_K 和 G_L 为新类 G_L 后类内离差平方和的增值。

由伪 t^2 统计量的定义知，值越大，表示 G_K 和 G_L 合并为新类 G_L 后类内离差平方和的增量 B_{KL}^2 相对于 G_K 和 G_L 的类内离差平方和越大，这表明上一次聚类的效果较好。

6.4.2 模糊聚类的有效性

模糊 C 均值聚类的需要预先给定分类数，如何确定最优的分类数，这就是聚类有效性所研究的内容。关于这方面的研究，现在依然是一个热点问题，至今为止，仍然没有一个最优的标准，只能是在相应的准则下最优。

对于二维数据，我们或许可以根据其平面图像大致看出分为几类合适，可是，对于高维数据，此方法就失效了，因此，有必要给出一些判别准则，比较有名的判别准则有：

1）Bezdek 提出的

$$V_{PE}=-\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}\ln(\mu_{ij})$$

其中 μ_{ij} 表示第 k 个数据点到第 i 类中心的隶属度，且 V_{PE} 的最小值点对应最佳聚类数。

2）Xie 和 Beni（XB）的准则

$$V_{XB}=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\parallel x_j-v_i\parallel^2}{n\min\limits_{i\neq j}\parallel v_j-v_i\parallel^2}$$

$\min\limits_{2\leqslant c\leqslant n-1}V_{XB}$ 对应的 c^* 即为最优聚类数。

3）Kuyama 和 Sugeno 的

$$V_{FS}=\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m\parallel x_j-v_i\parallel^2-\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^m\parallel v_i-\overline{v}\parallel^2$$

其中 $\overline{v}=\sum_{j=1}^{n}x_j$，$\mu_{ij}$ 表示第 j 个数据点到第 i 类中心的隶属度，V_{FS} 的最小值点对应最佳聚类数。

4）Kwon 的

$$V_k=\frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^2\parallel x_j-v_i\parallel^2+\frac{1}{c}\sum_{i=1}^{c}\parallel v_j-\overline{v}\parallel^2}{\min\limits_{i\neq k}\parallel v_i-v_k\parallel^2}$$

其中 μ_{ij} 表示第 j 个数据点到第 i 类中心的隶属度，V_k 的最小值点即为最佳聚类数。

例 6.4.3 对经典的 Iris 数据和葡萄酒数据，分别应用上述准则，确定最佳聚类数。

解： 利用 MATLAB 软件我们可以求得相应的最佳聚类数，见表 6-3。

表 6-3 两类经典数据 FCM 的最佳聚类数

m	准则函数	花蕾聚类数	葡萄酒聚类数	m	准则函数	花蕾聚类数	葡萄酒聚类数
1.5	V_{PE}	2	2	2.1	V_{PE}	2	2
	V_{XB}	2	2		V_{XB}	2	2
	V_{FS}	5	7		V_{FS}	5	6
	V_K	2	2		V_K	2	2
2	V_{PE}	2	2	2.5	V_{PE}	2	2
	V_{XB}	2	2		V_{XB}	2	2
	V_{FS}	5	11		V_{FS}	5	4
	V_K	2	2		V_K	2	2

习 题 6

1. 安徽省2008年各地市的森林资源见表6-4，求解以下问题：

表6-4 安徽省各市森林资源情况（2008年）

地 区	林业用地面积（千公顷）	森林面积（千公顷）	森林覆盖率（%）	活立木总蓄积量（万立方米）	森林蓄积量（万立方米）
合肥市	53.93	50.98	15.48	256.00	65.41
淮北市	44.92	40.38	14.99	211.07	151.14
亳州市	148.19	145.54	17.10	842.09	677.52
宿州市	293.86	279.86	28.80	1 238.01	1 035.67
蚌埠市	86.96	74.64	12.91	302.67	299.32
阜阳市	165.62	160.25	16.46	898.76	800.96
淮南市	17.93	16.37	6.20	151.39	30.17
滁州市	199.46	158.24	11.90	885.16	591.17
六安市	660.36	607.16	34.74	2 278.37	1 984.36
马鞍山市	17.14	13.72	8.10	81.20	36.34
巢湖市	148.52	117.54	12.60	494.38	335.26
芜湖市	77.27	66.69	20.85	279.34	187.92
宣城市	724.30	640.15	54.00	2 446.98	2 323.04
铜陵市	36.78	32.10	32.12	137.64	115.10
池州市	539.49	458.66	56.86	2 277.00	2 237.43
安庆市	598.92	546.67	35.60	2 291.09	2 099.21
黄山市	791.50	680.96	77.80	3 298.56	3 252.88

资料来源：《安徽统计年鉴2009》。

（1）在进行谱系聚类时，选择合适的类间距离，进而确定最优分类数，作出谱系聚类图。

（2）在进行模糊C均值聚类时，确定最优的分类数，并分析所得的结果。

（3）比较谱系聚类和模糊C均值聚类的结果，看看有什么异同。

2. 按来源分，2008年我国34个地区中的29个地区的农村居民家庭人均纯收入见表6-5，试利用K均值聚类和模糊减法聚类法进行聚类。

表6-5 农村居民人均家庭收入 （单位：元/人）

省 （市）	工资性收入	家庭经营纯收	财产性收入	转移性收入
北 京	6 389.31	2 058.57	1 142.8	1 071.25
河 北	4 064.95	3 097.14	463.39	285.3
山 西	1 979.52	2 416.22	118.63	281.09
内蒙古	1 713.55	1 986.38	153.05	244.26
辽 宁	806.48	3 218.01	114.9	516.79
黑龙江	2 035.53	2 931.26	201.29	408.4
上 海	810.17	3 344.72	183.2	594.66
江 苏	916.76	3 163.7	243.57	531.57
浙 江	8 108.32	711.26	849.83	1 770.85
安 徽	3 895.5	2 812	253.47	395.5
福 建	4 587.44	3 762.93	437.52	470.04

（续）

省　（市）	工资性收入	家庭经营纯收	财产性收入	转移性收入
江　西	1 737.84	2 114.24	119.04	231.37
山　东	2 421.46	3 146.09	179.03	449.49
河　南	1 842.36	2 552.59	66.55	235.69
湖　北	2 263.46	2 962.96	163.93	251.07
湖　南	1 499.93	2 699.3	53	202.02
广　东	1 742.33	2 690.83	40.82	182.4
广　西	1 990.52	2 196.61	57.06	268.26
海　南	3 684.47	2 001.5	339.47	374.35
重　庆	1 283.39	2 190.62	41.76	174.58
四　川	808.63	3 235.09	53.58	292.68
贵　州	1 764.64	2 016.64	50.9	294.03
云　南	1 620.4	2 061.7	71.37	367.74
西　藏	1 002.68	1 512.47	63.92	217.86
陕　西	617.47	2 156.8	109.83	218.5
甘　肃	759.72	1 845.04	185.46	385.6
青　海	1 243.57	1 475.01	86.01	331.87
宁　夏	867.98	1 543.24	19.49	293.08
新　疆	983.16	1 602.74	148.55	326.8

资料来源：《2009 中国统计年鉴》。

实验 5　聚类方法与聚类有效性

实验目的

1. 熟练掌握应用 MATLAB 软件计算谱系聚类与 K 均值聚类的命令。
2. 熟练掌握模糊 C 均值聚类与模糊减法聚类的 MATLAB 实现。
3. 掌握最优聚类数的理论及其实现。

实验数据与内容

2008 年我国 34 个地区中的 29 个地区的城镇居民人均收入见表 6-6。解决以下问题：

表 6-6　城镇居民人均收入　　（单位：元/人）

省（区、市）	工 薪 收 入	经营净收入	财产性收入	转移性收入
北　京	18 738.96	778.36	452.75	7 707.87
河　北	8 891.5	1 078.67	224.86	3 946.39
山　西	9 019.35	983.21	202.31	3 654.11
内蒙古	10 284.43	1 555.31	324.64	3 031.05
辽　宁	9 494.59	1 483.3	248.04	4 610.32
黑龙江	7 393.39	1 241.37	122.83	3 506.48
上　海	21 791.11	1 399.14	369.12	6 199.77

（续）

省（区、市）	工薪收入	经营净收入	财产性收入	转移性收入
江　苏	12 319.86	1 999.61	307.31	5 548.78
浙　江	15 538.83	3 161.87	1 324.94	4 955.14
安　徽	9 302.38	959.43	293.92	3 603.72
福　建	12 668.82	2 185.13	952.91	3 879.29
江　西	9 105.96	1 106.31	265.35	2 985.96
山　东	12 940.62	1 194.4	346.9	3 067.05
河　南	9 043.52	1 161.96	156.46	3 545.86
湖　北	9 474.81	1 114.68	244.13	3 340.65
湖　南	9 070.97	1 575.08	316.48	3 614.74
广　东	15 188.39	2 405.92	701.25	3 382.95
广　西	10 321.2	1 314.4	441.15	3 316.44
海　南	8 999.75	1 311.38	396.89	2 890.59
重　庆	10 957.62	788.26	205.94	3 265.92
四　川	9 117	1 040.14	262.9	3 265.06
贵　州	7 811.16	770.86	110.9	3 492.7
云　南	8 596.88	1 165.96	849.45	3 505.74
西　藏	12 314.69	303.34	138.08	891.42
陕　西	9 794.82	544	151.46	3 356.85
甘　肃	8 354.63	638.76	65.33	2 610.61
青　海	8 595.48	763.07	50.17	3 458.63
宁　夏	8 793.54	1 856.94	182.67	3 285.49
新　疆	9 422.22	938.15	141.75	1 976.49

资料来源：《中国统计年鉴 2009》。

（1）计算各样品间的欧氏距离、马氏距离和加权平方距离。

（2）运用谱系聚类法进行聚类，包括确定最优聚类数，选择合适的类间距离，同时作出谱系图。

（3）运用 K 均值聚类法进行聚类。

（4）运用模糊 C 均值聚类和模糊减法聚类法进行聚类。

（5）综合分析以上不同的聚类法所得的聚类结果，能得到什么样的结论？

第7章 数值模拟分析

数值模拟以计算机为手段，通过数值计算和图像显示的方法，达到对工程数学问题和物理问题乃至自然界各类问题研究的目的。数值模拟技术在电子信号、图像识别、金融数据分析等领域有着广泛的应用。本章主要介绍随机数值的模拟方法以及利用BP神经网络进行模式识别与预测的方法。

7.1 蒙特卡罗方法与应用

7.1.1 蒙特卡罗方法的基本思想

蒙特卡罗方法又称统计试验方法，它是一种采用统计抽样理论近似求解数学问题和物理问题的方法。它既可以用来研究概率问题，也可用来求解非概率问题。为使读者了解为什么可以用概率统计的方法来解决数学计算问题，从而抓住蒙特卡罗方法的思想实质，我们着重从求解非概率问题中选取一些简单而富有启发性的例子加以说明。

利用蒙特卡罗方法解决数学分析问题，基本的想法是首先建立与描述该问题有相似性的概率模型。利用这种相似性把概率模型的某些特征（如随机事件的概率或随机变量的平均值等）与数学问题的解答（如积分值等）联系起来，然后对模型进行随机模拟或统计抽样，再利用所得结果求出这些特征的统计估计值作为原来的分析问题的近似解。

例如，考虑积分

$$I=\int_0^1 f(x)\mathrm{d}x$$

假设当$0\leqslant x\leqslant 1$时$0\leqslant f(x)\leqslant 1$，这时积分$I$等于由曲线$y=f(x)$、$Ox$轴、$Oy$轴以及直线$x=1$围成的区域$G$的面积（如图7-1所示）。为求此面积，我们设想在正方形$\{0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant 1\}$内随机地投掷一个点，落点的两个坐标是相互独立且在区间（0，1）上的均匀分布（即每点都具有等可能性）。那么这个落点在曲线$y=f(x)$以下的区域G内的概率p显然等于这区域的面积。以（X，Y）表示正方形内的任一点的坐标。如果我们用某种方法得到均匀分布的独立变量X及Y的n个样本值，对（X，Y）的每一取样值$(x_i,y_i)(i=1,2,\cdots,n)$检查$y_i$与$f(x_i)$是否满足不等式

$$y_i<f(x_i) \tag{7.1.1}$$

如果（7.1.1）式成立，说明点（x_i，y_i）落在区域G内（如图7-1中的点（x_1，y_1）），否则落在区域G外（如图7-1中的点（x_2，y_2））。设满足不等式（7.1.1）的点数为m，则由大数定律知，当n足够大时频率近

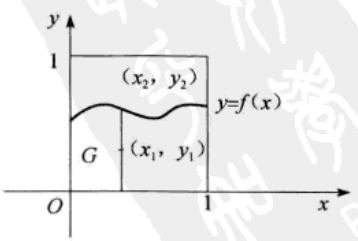


图7-1 区域G示意图

似于点落在区域 G 内的概率 p，即

$$I=\int_0^1 f(x)\mathrm{d}x\approx\frac{m}{n} \tag{7.1.2}$$

我们还可以利用随机变量的平均值（数学期望）来计算积分 I。设随机变量 X 在区间（0，1）上服从均匀分布，$Y=f(X)$，则由期望的计算公式知

$$E(Y)=E(f(X))=\int_0^1 f(x)\cdot 1\mathrm{d}x=I$$

如果我们用某种方法得到均匀分布的变量 X 的 n 个样本值 $x_i(i=1,2,\cdots,n)$，计算变量 Y 的 n 个样本值 $y_i=f(x_i)$，则由大数定律知

$$E(Y)\approx\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{1}{n}\sum_{i=1}^{n}f(x_i)$$

即

$$I=\int_0^1 f(x)\mathrm{d}x\approx\frac{1}{n}\sum_{i=1}^{n}f(x_i) \tag{7.1.3}$$

由以上求解过程可以看出，当所求问题的解是某个事件的概率，或者是某个随机变量的数学期望，或者是与概率、数学期望有关的量时，通过某种试验的方法，得出该事件发生的频率，或者该随机变量若干个具体观察值的算术平均值，通过它得到问题的近似解。

7.1.2　随机数的产生与 MATLAB 的伪随机数

在上面的积分计算中，要求获得在（0，1）上均匀分布的随机变量 X 的一系列取样值。一般地说，利用蒙特卡罗方法作各种类型数值计算时，同样也必须找出模拟随机变量或随机过程的现实，为此就需要所谓的随机数。最常用的随机数是在（0，1）上均匀分布的随机数，可以证明任意其他分布律的随机数可以利用均匀分布的随机数来产生。随机数的生成通常有两种方法：一种是依赖一些专用的电子元件发出随机信号，这种方法又称为物理生成法；另一种是通过数学的算法，仿照随机数发生的规律计算出随机数。目前，许多计算机系统都有随机数生成函数，调用它们便可生成需要的随机数。计算程序产生的随机数不是真正的随机数，它们是确定的，但看上去是随机的，且能通过一些随机性的检验，故常称为伪随机数。表 7-1 列出了 MATLAB 的随机数的生成函数及使用说明。

表 7-1　常见分布随机数的生成函数

随机数名称	命令调用格式	参 数 说 明
(0，1) 上均匀分布	Y=rand(m，n)	生成区间（0，1）上的均匀分布随机数
二项分布	Y=binornd(N，p，m，n)	生成参数为 N、p 的 m 行 n 列的 $m\times n$ 个二项分布随机数
几何分布	Y=geornd(p，m，n)	生成参数为 p 的 m 行 n 列的 $m\times n$ 个几何分布随机数
泊松分布	Y=poissrnd(λ，m，n)	生成参数为 λ 的 m 行 n 列的 $m\times n$ 个泊松分布随机数
均匀分布	Y=unifrnd(a，b，m，n)	生成区间（a、b）上的 m 行 n 列的 $m\times n$ 个均匀分布随机数
指数分布	Y=exprnd(λ，m，n)	生成参数为 λ 的 m 行 n 列的 $m\times n$ 个指数分布随机数
正态分布	Y=normrnd(μ，σ，m，n)	生成参数为 μ、σ 的 m 行 n 列的 $m\times n$ 个正态分布随机数
T 分布	Y=trnd(k，m，n)	生成自由度为 k 的 m 行 n 列的 $m\times n$ 个 T 分布随机数
χ^2 分布	Y=chi2rnd(k，m，n)	生成自由度为 k 的 m 行 n 列的 $m\times n$ 个 χ^2 分布随机数
对数正态分布	R=lognrnd(μ，σ，m，n)	生成参数为 μ、σ 的 $m\times n$ 个对数正态分布随机数
Beta 分布	R=betarnd(A，B，m，n)	生成参数为 A、B 的 $m\times n$ 个 Beta 分布随机数

7.1.3　蒙特卡罗方法应用实例

1. 圆周率的模拟

例 7.1.1　用蒙特卡罗方法模拟求圆周率 π 的近似值。

解：（频率法）设二维随机变量（X，Y）在正方形 $\{0\leqslant x\leqslant 1,\ 0\leqslant y\leqslant 1\}$ 内服从均匀分布，如图 7-2 所示，则（X，Y）落在圆内的概率为

$$p = P\{X^2+Y^2\leqslant 1\} = \frac{\pi}{4}$$

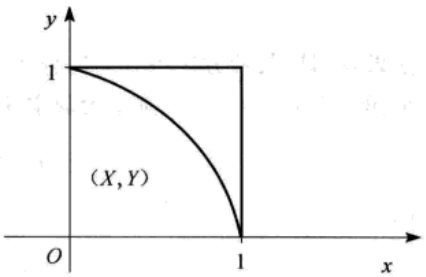


图 7-2　正方形区域

产生均匀分布的独立变量 X 及 Y 的 n 个样本值，对（X，Y）的每一取样值（x_i，y_i）($i=1,2,\cdots,n$）检查随机数是否满足：$x_i^2+y_i^2\leqslant 1$（相当于第 i 个随机点落在 1/4 圆内）。若有 k 个点落在 1/4 圆内，则随机事件“点落入 1/4 圆内”的频率为 k/n。根据大数定律，$p=\frac{\pi}{4}\approx\frac{k}{n}$，所以

$$\pi\approx\frac{4k}{n}$$

在 MATLAB 中编程模拟计算得：$\pi\approx 3.1417$。

程序如下：

```
k= 0;                                  % k 用于随机点落在 1/4 圆内的计数
for j= 1:100000                        % 样本个数取为 N= 100000
a= rand(1,2);                          % 生成区间(0,1)上的均匀分布随机数作取样值
if a(1)^2+ a(2)^2< = 1                 % 检查随机数是否满足:x1^2+ y1^2≤1
k= k+ 1;
end
end
PI= 4 * k/j                            % 计算 π 的近似值
```

注意　由于是模拟计算，所以当读者将程序再运行时所得的结果不一定是 3.1417。

（平均值法）因为，$\int_0^1\sqrt{1-x^2}\,\mathrm{d}x=\frac{\pi}{4}$ 令

$$f(x)=\sqrt{1-x^2}$$

所以当随机变量 X 在区间 [0，1] 上服从均匀分布时，

$$\frac{\pi}{4}=\int_0^1\sqrt{1-x^2}\,\mathrm{d}x=E(f(X))$$

产生 X 的均匀分布的 n 个样本值 $x_i(i=1,2,\cdots,n)$，由（7.1.3）式

$$\pi=4E(f(X))\approx\frac{4}{n}\sum_{i=1}^{n}\sqrt{1-x_i^2}$$

在 MATLAB 中编程模拟计算得：$\pi\approx 3.1413$。

程序如下：

```
x= rand(1,100000);          % 生成区间(0,1)上的均匀分布随机数 100000 个
y= sqrt(1- x.^2);           % 计算 f(xi)= √(1-xi^2)
PI= 4 * mean(y)             % 计算 π 的近似值
```

从模拟的结果看，模拟值与真实值很接近。

2. 资产价格的模拟

例 7.1.2　（股票价格变化的模拟）假设股票在 t（单位：天）时刻的价格为 $S(t)$（单位：元），且满足随机微分方程

$$dS(t) = S(t)[\mu dt + \sigma dZ(t)] \tag{7.1.4}$$

其中 $dZ(t)=\varepsilon\sqrt{dt}$，$Z(t)$是维纳过程或称布朗运动（Brownian motion），$\varepsilon\sim N(0,1)$；μ 为股票价格的期望收益率；σ 为股票价格的波动率。又假设股票在 $t=t_0$ 时刻的价格为 $S_0=S(t_0)=20$，期望收益率为 $\mu=0.031$（单位：元/年），波动率 $\sigma=0.6$，试用蒙特卡罗方法模拟未来 90 天的价格曲线，并确定未来第 90 天股票价格的分布图。

解：MATLAB 脚本程序如下：

```
dt= 1/365.0;                          % 一天的年单位时间
S0= 20;                               % 股票在初始时刻的价格,程序中假设
r= 0.031;                             % 期望收益率
sigma= 0.6;                           % 波动率 σ= 0.6
expTerm= r * dt;                      % 漂移项 μdt
stddev= sigma * sqrt(dt);             % 波动项 σdz(t)
nDays1= 90;                           % 要模拟的总天数
for nDays= 1: nDays1                  % nDays 表示时刻 t
nTrials= 10000;                       % 模拟次数
for j= 1:nTrials
n =  randn(1,nDays);                  % 生成 nDays 个标准正态分布随机数
S= S0;
for i= 1:nDays
dS =  S * (expTerm+ stddev * n(i));   % 模拟计算股票价格的增量
S= S+ dS;                             % 计算股票价格
end
S1(nDays,j)= S;                       % 将每天的股票模拟价格数据记录在 S1 中
end
end
S2= mean(S1');                        % 计算每天模拟的股票价格的均值,作为价格的估值
plot(S2','- o')                       % 90 天期间股票价格估值的曲线图
hist(S1(90,:)(,0:.5:65)               % 第 90 天的股票价格模拟的直方图
```

股票未来 90 天的价格走势图，如图 7-3 所示。股票第 90 天的价格模拟图，如图 7-4 所示。

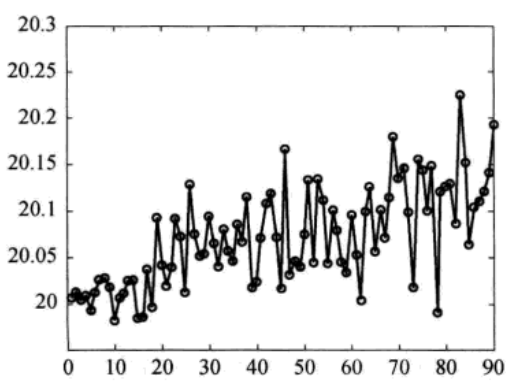


图 7-3　股票未来 90 天的价格走势图

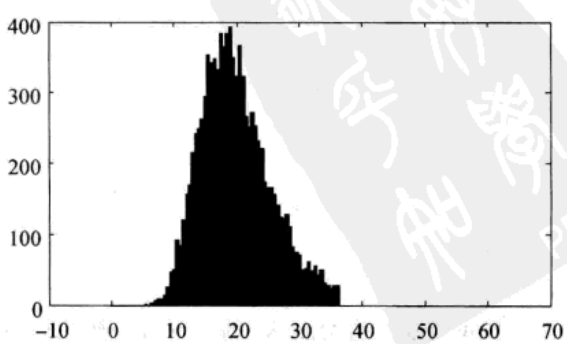


图 7-4　股票第 90 天的价格模拟图

3. 马尔科夫链的模拟

例 7.1.3 （稳态马尔科夫链的模拟）考虑 101×101 个点构成的正方形区域定义其上的整数点组成的集合

$$A = \{(i,j) \mid i = 1,2,\cdots,100; j = 1,2,\cdots,100\}$$

选取初始点 $X_0=(10, 80)$，按照如下步骤生成马尔科夫链：（1）选择基准点（0，0），(100，0)，(50，100)；（2）随机等可能地选择一个基准点，计算初始点与该基准点的中点坐标并作为新的初始点；（3）重复步骤（1）、(2)。基准点的集合记为

$$R = \{(0,0),(100,0),(50,100)\}$$

记 $\{r_n \mid n = 0,1,2,\cdots\}$ 表示第 n 步选择的基准点，则 $\{r_n\}$ 的状态空间为 R，其一步转移概率为

$$p_{ij} = P\{r_{n+1} = r_j \mid r_n = r_i\} = \frac{1}{3} \tag{7.1.5}$$

即 $\{r_n\}$ 是马尔科夫链。又设 $\{X_n=(x_n, y_n) \mid (x_n, y_n)\in A\}$ 表示第 n 步的初始点位置 $(n = 0,1,2,3,\cdots)$。要求：（1）模拟 $X_n=(x_n, y_n)$ 的轨迹；（2）改变初始点 X_0 的位置，观察轨迹的变化；（3）改变基准点的位置，观察轨迹的变化。

解：依题意，初始点位置与基准点满足

$$X_{n+1} = \frac{1}{2}(X_n + r_n)(n = 0,1,2,\cdots) \tag{7.1.6}$$

其中 $\{r_n\}$ 是马尔科夫链。

编写 MATLAB 程序文件如下：

```
clear all
rand('state',0)                                  % 返回生成随机数的初始状态
R= [0,100,50;0,0,100];                           % 基准点坐标或状态空间
x0= [10 80]';                                    % 设置初始点
plot(x0(1),x0(2),'.')                            % 输出初始点图形
axis([0 100 0 100])                              % 纵横坐标范围
hold on
xn_1= x0;
for n= 1:10000                                   % 选择初始点的次数为 10000 次
      j= floor(3* rand(1,1)+ 1);                 % 随机选择三个基准点中的一个
      xn(:,n)= round(0.5* (R(:,j)+ xn_1));       % 按(7.1.6)式生成新的初始点
      plot(xn(1,:),xn(2,:),'.')                  % 输出新的初始点图形
      xn_1= xn(:,n);
end
grid
hold off
```

马尔科夫链的初始点的轨迹如图 7-5 所示。

从图形可以看到：初始点的轨迹分布在以三个基准点为顶点的一个正三角形区域内，形成一个稳定的图形结构，且初始点在正方形区域内对有些点永远到达不了。

当我们改变初始点 X_0 的位置时，初始点的轨迹分布仍在同一个正三角形区域内。当我们改变基准点时，初始点的轨迹分布在以三个基准点为顶点的三角形区域内。这一现象的模拟

结果可以从理论上进行逻辑证明（此处略）。

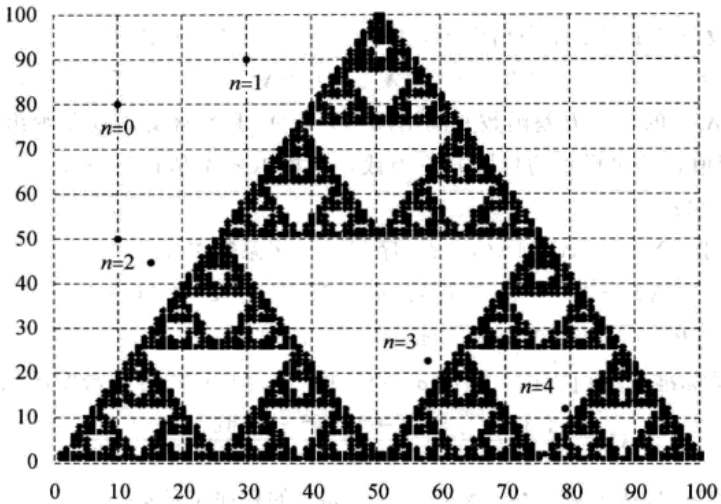


图 7-5　马尔科夫链的初始点的轨迹

4. 时间序列模拟

设 $\{\varepsilon_t\}$ 是白噪声 $WN(0,\sigma^2)$，实系数多项式 $A(z)$ 和 $B(z)$ 没有公共根，满足 $b_0=1$，$a_p b_q \neq 0$，$A(z)=1-\sum_{j=1}^{p} a_j z^j \neq 0(|z|\leqslant 1)$，$B(z)=\sum_{j=0}^{q} b_j z^j \neq 0(|z|<1)$，我们称差分方程

$$X_t=\sum_{j=1}^{p} a_j X_{t-j}+\sum_{j=0}^{q} b_j \varepsilon_{t-j} \quad (t\in \mathrm{Z})$$

是一个自回归滑动模型，简称为 ARMA(p，q) 模型。称平稳序列 $\{X_t\}$ 为 ARMA(p，q) 序列。其中 p、q 是正整数，Z 是整数集合。

ARMA(p，q) 序列 $\{X_t\}$ 的自协方差函数可以由 Wold 系数 $\{\psi_j\}$ 表示：

$$\gamma_k=\sigma^2\sum_{j=0}^{\infty}\psi_j\psi_{j+k} \quad (k\geqslant 0) \tag{7.1.7}$$

其中系数 $\{\psi_j\}$ 采用如下的递推方法：

$$\psi_j=\begin{cases}1 & j=0\\ b_j+\sum_{k=1}^{p} a_k\psi_{j-k} & j=1,2,\cdots\end{cases}$$

且规定：当 $j>q$ 时，$b_j=0$；当 $j<0$ 时，$\psi_j=0$

ARMA(p，q) 序列有谱密度函数：

$$f(\lambda)=\frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\gamma_k \mathrm{e}^{-\mathrm{i}k\lambda}=\frac{\sigma^2}{2\pi}\left|\frac{B(\mathrm{e}^{\mathrm{i}\lambda})}{A(\mathrm{e}^{\mathrm{i}\lambda})}\right|^2 \tag{7.1.8}$$

当平稳序列 $\{X_t\}$ 的 N 个样本观测值为

$$x_1,x_2,\cdots,x_N$$

时，序列 $\{X_t\}$ 的样本自协方差为

$$\hat{\gamma}_k=\frac{1}{N}\sum_{j=1}^{N-k}(x_j-\overline{x}_N)(x_{j+k}-\overline{x}_N) \quad (0\leqslant k\leqslant N-1) \tag{7.1.9}$$

它是 $\{X_t\}$ 的自协方差函数 $\gamma_k=\mathrm{cov}(X_1, X_{k+1})$ 的点估计，且 $\{X_t\}$ 的谱密度估计为

$$\hat{f}(\lambda)=\frac{1}{2\pi}\sum_{k=-\infty}^{\infty}\hat{\gamma}_k e^{-ik\lambda} \tag{7.1.10}$$

例 7.1.4 设$\{\varepsilon_t\}$是标准正态白噪声，序列$\{X_t,\ t\in\mathbf{Z}\}$满足

$$X_t=0.9X_{t-1}+1.4X_{t-2}+0.7X_{t-3}+0.6X_{t-4}+\varepsilon_t+0.5\varepsilon_{t-1}-0.4\varepsilon_{t-2} \tag{7.1.11}$$

(1) 计算$\{X_t\}$的自协方差函数并画出图形；(2) 求谱密度函数并画出图形；(3) 模拟X_t的300个观测值，求出样本的自协方差函数，由模拟样本估计谱密度。将估计谱密度与真实谱密度函数作比较。

解：容易验证$\{X_t\}$是ARMA(4，2)序列，且实系数多项式

$$A(z)=1-0.9z+1.4z^2+0.7z^3+0.6z^4\neq 0(|z|\leqslant 1)$$

$$B(z)=1+0.5z-0.4z^2\neq 0(|z|<1)$$

由于$\{\varepsilon_t\}$是标准正态白噪声，因此$\sigma^2=1$。由(7.1.8)式得谱密度函数

$$f(\lambda)=\frac{1}{2\pi}\left|\frac{1+0.5e^{i\lambda}-0.4e^{2i\lambda}}{1-0.9e^{i\lambda}+1.4e^{2i\lambda}+0.7e^{3i\lambda}+0.6e^{4i\lambda}}\right|^2$$

(1) 计算$\{X_t\}$自协方差函数$\{\gamma_k:\ 0\leqslant k\leqslant 23\}$的程序文件如下。

建立M文件：litit7_1_4.m

```
a=[-0.9,-1.4,-0.7,-0.6];                    % 输入自回归部分的系数
b=[0.5,-0.4];                               % 输入移动部分的系数
% 计算Wold系数
c(1)=1;                                     % 用c数组记录Wold系数
c(2)=b(1)+a(1);
c(3)=b(2)+a(1)*c(2)+a(2)*c(1);
c(4)=a(1)*c(3)+a(2)*c(2)+a(3)*c(1);
c(5)=a(1)*c(4)+a(2)*c(3)+a(3)*c(2)+a(4)*c(1);
for k=1:1000
c(5+k)=a(1)*c(4+k)+a(2)*c(3+k)+a(3)*c(2+k)+a(4)*c(1+k);
end
% 计算理论自协方差函数
for m=0:23                                  % m表示自协方差函数的自变量
    h=0;
    for k=1:980
        h=h+c(k)*c(k+m);                    % 计算自协方差函数
    end
    d(m+1)=h;                               % 向量d表示自协方差函数的值向量
end
d                                           % 输出自协方差函数值
plot(d)                                     % 绘制自协方差函数图形
grid on
```

输出$\{X_t\}$的前24个自协方差函数$\{\gamma_k:\ 0\leqslant k\leqslant 23\}$如下(横读)：

```
  6.6708    -1.5078    -4.5792     2.4672     1.2433    -0.4630    -0.3035    -1.4293
  1.2894     1.3309    -1.8203    -0.2699     1.0861    -0.1239    -0.1279    -0.3097
 -0.1071     0.6939    -0.1810    -0.5477     0.3249     0.1848    -0.1292    -0.0412
```

序列(7.1.11)的自协方差函数图形如图7-6所示。

（2）谱密度函数图形程序：

```
b= [0.5,- 0.4];
sgm= 1;
a= [- 0.9,- 1.4,- 0.7,- 0.6];
t= 0:0.02:pi;
a1= abs(1+ b(1) * exp(i * t)+ b(2) * exp(2* i * t));    % (7.1.8)式中的| A(z)|
a2= abs(1- a(1) * exp(i * t)- a(2) * exp(2 * i * t)- a(3) * exp(3 * i * t)- a(4) * exp(4 * i *
    t));                                                % | B(e^iλ)|
ft= sgm* (a1./a2).^2/2/pi;                              % (7.1.8)式
plot(t,ft)                                              % 绘图
grid on
```

序列（7.1.11）的谱密度函数图形如图 7-7 所示。

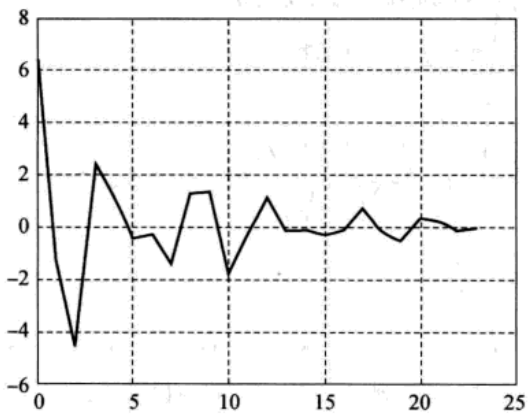


图 7-6　序列（7.1.11）的自协方差函数图形

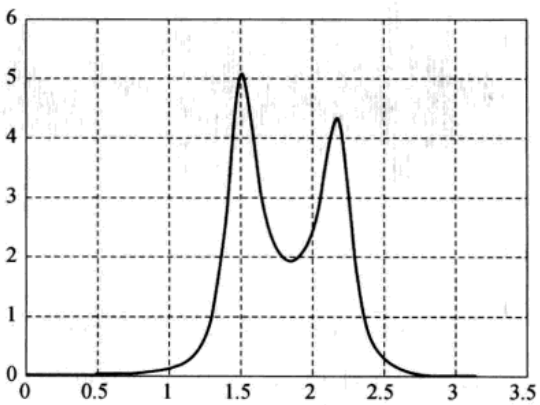


图 7-7　序列（7.1.11）的谱密度函数图形

谱密度的图形有两个峰值，说明这个平稳序列有两个频率成分。

（3）模拟序列（7.1.11）的 300 个观测值，并求出样本自协方差函数与谱密度函数的估计。编写程序并保存成名为“liti7_1_4_3.m”的 M 文件程序如下：

```
% 模拟样本 liti7_1_4_3
a= [- 0.9,- 1.4,- 0.7,- 0.6];      % 输入自回归部分的系数
b= [0.5,- 0.4];                    % 输入移动部分的系数
for h1= 1:1000                     % 模拟生成序列(7.1.11)的数据 1000 次
   xt= randn(1,364);               % 模拟白噪声过程的 364 个观测值
   yt(h1,1)= 0;  yt(h1,2)= 0; yt(h1,3)= 0; yt(h1,4)= 0;   % 给序列赋初值
   for k= 1:360
      yt(h1,k+ 4)= sum(a.* yt(h1,[k+ 3:- 1:k]))+ sum([1,b].* xt([k+ 4:- 1:k+ 2]));
                                                          % 模拟序列数据
   end
     rt1(h1,:)= autocorr(yt(h1,[65:364]),24);             % 自相关系数
     rt2(h1,:)= std(yt(h1,[65:364]))^2 * rt1(h1,:);       % 自协方差函数
end
figure
myt= mean(yt(:,[65:364]));                                % 1000 次模拟的平均值作为序列的样本值
plot(myt)                                                 % 样本散点图
```

```
figure
rt= mean(rt2);                                    % 1000 次模拟样本值的自协方差函数平均值
plot(rt([1:24]))                                  % 绘制模拟样本自协方差函数图形
gtext('模拟样本自协方差函数')
grid on
% 在命令窗口中输入以下命令用于作出理论自协方差函数图形
hold on                                           % 保持图形窗口
liti7_1_4                                         % 调用文件 liti7_1_4.m 绘制理论自协方差函数图形
```

程序运行输出结果如图 7-8 和图 7-9 所示。

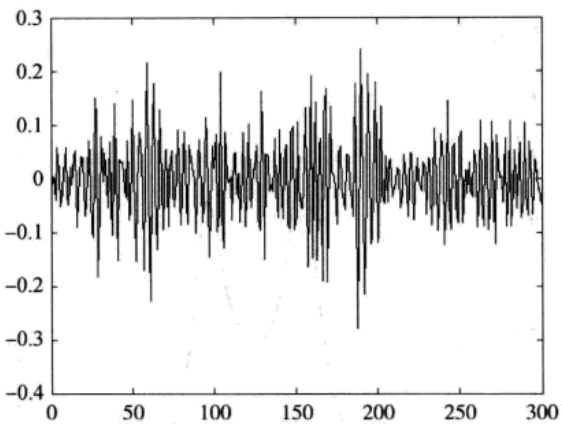


图 7-8 序列（7.1.11）的 300 个模拟值
（ave=−9.881 8e−004，std=0.081 0）

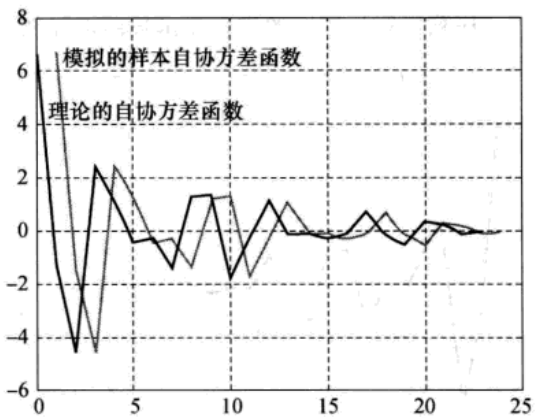


图 7-9 序列（7.1.11）理论与模拟样本自协方差函数图形

注意 由于模拟的自协方差函数曲线与理论自协方差函数曲线几乎重合，图 7-9 中模拟的样本自协方差函数曲线向右平移了 1 个单位。

在 MATLAB 命令窗口中打开 liti7_1_4_3.m 文件，并输入如下程序：

```
% 谱密度的估计
liti7_1_4_3                                       % 打开 liti7_1_4_3.m 文件
t= 0:0.02:pi;                                     % 谱密度的自变量取值
for j= 1:length(t)
f(j)= sum(rt([1:25]).*exp(i*(0:24)*t(j)))+ sum(rt([2:25]).*exp(- i*(1:24)*t(j)));
f(j)= f(j)/2/pi;                                  % 由(7.1.10)计算谱密度的估计值
end
figure
plot(t,f,'r')                                     % 绘模拟样本的谱密度的图形
grid on
% 在命令窗口中输入以下命令,用于理论谱密度与模拟样本的谱密度的图形比较
hold on
liti7_1_4_2                                       % 调用理论谱密度的图形
```

输出谱密度函数的估计图形如图 7-10 所示。

图 7-10 还同时画出了理论谱密度，从图形上看，理论谱密度与估计的谱密度非常接近。这说明用 MATLAB 进行模拟分析往往会取得很好的效果。

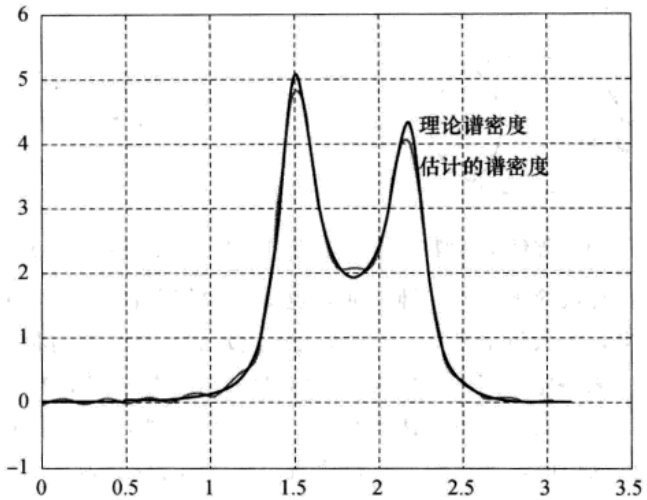


图 7-10　序列（7.1.11）模拟的样本估计的谱密度函数图形

7.2　BP 神经网络及应用

7.2.1　人工神经元及人工神经元网络

人工神经元网络是由大量的人工神经元经广泛互连形成的人工网络，用以模拟人类神经系统的结构和功能。

1. 人工神经元的结构

人工神经元是对生物神经元的抽象与模拟。所谓抽象是从数学角度而言的，所谓模拟是从其结构和功能而言的。1943 年，心理学家麦克洛奇（W.McMulloch）和数理逻辑学家皮茨（W.Pitts）根据生物神经元的功能和结构，提出了一个将神经元看作二进制阈值元件的简单模型，即 M-P 模型，如图 7-11 所示。

在图 7-11 中，x_1，x_2，…，x_n 表示某一神经元的 n 个输入；w_i 表示第 i 个输入的连接强度，称为连接权值；θ 为神经元的阈值；y 为神经元的输出。可以看出，人工神经元是一个具有多输入，单输出的非线性器件。它的输入为

$$\sum_{i=1}^{n}\omega_i x_i \tag{7.2.1}$$

输出为

$$y = f(\sigma) = f\Big(\sum_{i=1}^{n}\omega_i x_i - \theta\Big) \tag{7.2.2}$$

其中，f 称为神经元功能函数或作用函数。

2. 常用的人工神经元模型

功能函数 f 是表示神经元输入与输出之间关系的函数，根据功能函数的不同，可以得到不同的神经元模型。常用的神经元模型有以下几种。

（1）阈值型（Threshold）

这种模型的神经元没有内部状态，功能函数 f 是一个阶跃函数，它表示激活值 σ 和输出之

间的关系，如图 7-12 所示。

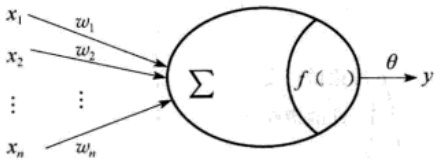


图 7-11　M-P 神经元模型

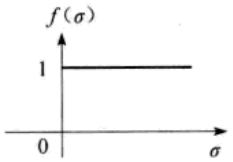


图 7-12　阈值型神经元的输入/输出特性

阈值型神经元是一种最简单的人工神经元，也就是我们前面提到的 M-P 模型。这是一种二值型神经元，其输出状态取值 1 或 0，分别代表神经元的兴奋和抑制状态。某一时刻，神经元的状态由功能函数 f 来决定。当激活值 $\sigma>0$ 时，即神经元输入的加权总和超过给定的阈值时，该神经元被激活，进入兴奋状态，其状态 $f(\sigma)$ 为 1；否则，当 $\sigma\leqslant 0$ 时，即神经元输入的加权总和不超过给定的阈值时，该神经元不被激活，其状态 $f(\sigma)$ 为 0。

（2）分段线性强饱和型（linear saturation）

这种模型又称为伪线性，其输入/输出之间在一定范围内满足线性关系，一直延续到输出为最大值 1 为止。但当达到最大值后，输出就不再增大。如图 7-13 所示。

（3）S 型（sibmoid）

这是一种连续的神经元模型，其输出函数也是一个有最大输出值的非线性函数，其输出值是在某个范围内连续取值的，输入输出特性常用指数、对数或双曲正切等 S 型函数表示。它反映的是神经元的饱和特性，如图 7-14 所示。

（4）子阈累积型（subthreshold summation）

这种类型的作用函数也是一个非线性函数，当产生的激活值超过 T 值时，该神经元被激活产生反响。在线性范围内，系统的反响是线性的，如图 7-15 所示。

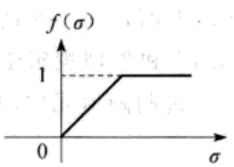


图 7-13　分段线性强饱和型神经元的输入/输出特性

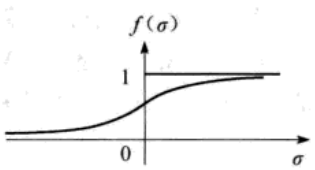


图 7-14　S 型神经元的输入/输出特性

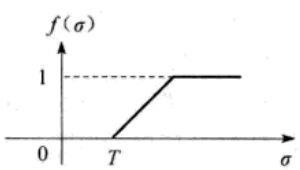


图 7-15　子阈累积型神经元的输入/输出特性

3. 人工神经网络

人工神经网络是一种应用类似于大脑神经突触联接的结构进行信息处理的数学模型。在工程与学术界也常直接简称为“神经网络”或“类神经网络”。神经网络是一种运算模型，由大量的节点（或称神经元、单元）和它们之间的相互连接构成。每个节点代表一种特定的输出函数，称为激励函数（activation function）。每两个节点间的连接都代表一个通过该连接信号的加权值，称之为权重（weight），这相当于人工神经网络的记忆。网络的输出则依网络的连接方式、权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。

7.2.2　BP 神经网络

在神经网络中，最具代表性和应用最广泛的是美国加州大学的鲁梅尔哈特（Rumelhart）

和麦克莱兰（Meclelland）等人于 1985 年提出的 BP（Bake-Propagation）神经网络（多层前馈式误差反向传播神经网络），该模型是一种有监督学习模型，具有很强的自组织、自适应能力的模型。它通过对有代表性的样本的学习训练后能掌握研究系统的本质特性，且结构简单、可操作性强，能模拟任意的非线性输入输出关系。

1. BP 神经网络的拓扑结构

从结构上看，BP 网络是典型的多层网络，它不仅有输入层节点、输出层节点，而且有一层或多层隐含节点。在 BP 网络中，层与层之间多采用全互连方式，但同一层的节点之间不存在相互连接。一个三层 BP 网络的结构如图 7-16 所示。

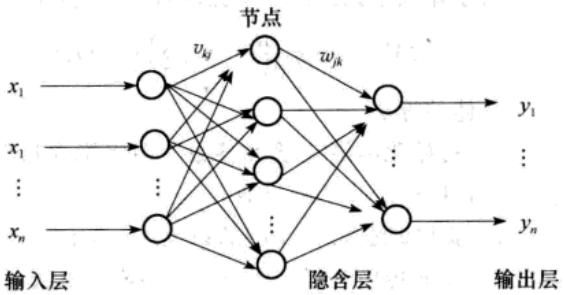


图 7-16　BP 网络的拓扑结构示意图

2. BP 神经网络模型权值问题的数学描述

假设取得 N 个样本 $\{y(t),\ x(t);\ t=1,2,\cdots,N\}$，其中 y 是 n 维向量，x 是 m 维向量，当第 t 个样本 $x(t)=(x_1(t),x_2(t),\cdots,x_m(t))$ 的数据输入网络时，网络相应的输出记为 $\hat{f}(t)=(\hat{f}_1(t),\hat{f}_2(t),\cdots,\hat{f}_n(t))$；隐含单元的状态记为 $H_k(t)(k=1,2,\cdots,q)$；从输入层到隐含层的权值记为 $V_{jk}(j=1,2,\cdots,m;k=1,2,\cdots,q)$；从隐含层到输出层之间的权值记为 $W_{ki}(k=1,2,\cdots,q;i=1,2,\cdots,n)$；隐含层的传递函数为 $g(\cdot)$，输出层的传递函数为 $f(\cdot)$，则隐含层节点的输出为（将阈值写入求和项中）：

$$H_k(t)=g\left(\sum_{j=0}^{m}V_{jk}x_j(t)\right)\quad(k=1,2,\cdots,q)\tag{7.2.3}$$

输出层节点的输出为：

$$\hat{f}_i(t)=f\left(\sum_{k=0}^{q}W_{ki}H_k(t)\right)=f\left(\sum_{k=0}^{q}W_{ki}g\left(\sum_{j=0}^{m}V_{jk}x_j(t)\right)\right)\quad(i=1,2,\cdots,n)\tag{7.2.4}$$

显然，对任何一组确定的输入，输出是所有权 $\{V_{jk},\ W_{ki}\}$ 的函数。确定权值的目标是确定适当的权 $w=\{V_{jk},\ W_{ki}\}$，使得网络的输出与系统的实际输出（或理想输出）误差最小。这归结为最优化问题，即确定适当的权 w，使

$$E(w)=\frac{1}{2}\sum_{i,t}(y_i(t)-\hat{y}_i(t))^2=\frac{1}{2}\sum_{i,t}\left[y_i(t)-f\left(\sum_{k=0}^{q}W_{ki}g\left(\sum_{j=0}^{m}V_{jk}x_j(t)\right)\right)\right]^2\tag{7.2.5}$$

达到极小。

对于权 w 来说，$E(w)$ 是一个连续可微的非线性函数，其极值一定存在，应用 BP 算法可以求解这一问题，下面介绍这一算法。

3. BP 算法原理

BP 算法具有梯度性，也称为最速下降法，是一种迭代算法，基本想法是：从一个初始点 w_0 出发，计算在 w_0 点的负梯度方向 $-\nabla E(w_0)$，只要 $\nabla E(w_0)\neq 0$，就可沿着该方向移动一小段距离，达到一个新的点 $w_1=w_0-\eta\nabla E(w_0)$，$\eta(\eta>0)$ 是参数，只要 η 足够小，定能保证 $E(w_1)<E(w_0)$。不断重复这一过程，一定能达到 $E(w)$ 的一个极小值。

对于神经网络来说，BP 算法由数据流的前向计算（正向传播）和误差信号的反向传播两个过程构成。正向传播时，传播方向为输入层→隐含层→输出层，每层神经元的状态只影响下一层神经元。若在输出层得不到期望的输出，则转向误差信号的反向传播流程。通过这两个过程的交替进行，在权向量空间执行误差函数梯度下降策略，动态迭代搜索一组权向量，

使网络误差函数达到最小值，从而完成信息提取和记忆过程。

对于隐含单元到输出单元的权 W_{ki}，最速下降法给出每一步的修正量是

$$\Delta W_{ki} = -\eta \frac{\partial E}{\partial W_{ki}} = \eta \sum_t \delta_i(t) H_k(t), \quad \text{其中 } \delta_i(t) = g(h_i(t))[y_i(t) - \hat{y}_i(t)]$$

对于输入单元到隐含单元的权 V_{jk}，修正量是

$$\Delta V_{jk} = -\eta \frac{\partial E}{\partial V_{jk}} = \eta \sum_t \delta_j(t) x_k(t), \quad \text{其中 } \delta_j(t) = g(h_j(t)) \sum_i \delta_i(t) W_{ki}$$

具体的算法：

1）初始化网络及学习参数，即将隐含层和输出层各节点的连接权值、神经元阈值赋予[－1，1]区间的一个随机数。

2）提供训练模式，即从训练模式集合中选出一个训练模式，将其输入模式和期望输出送入网络。

3）正向传播过程，即对给定的输入模式，从第一隐含层开始，计算网络的输出模式，并把得到的输入模式与期望模式进行比较，若有误差，则执行第 4）步，否则，返回第 2）步，提供下一个训练模式。

4）反向传播过程，即从输出层反向计算到第一隐含层，按以下方式逐层修正各单元的连接权值：

①计算同一层单元的误差 δ_k。

②按下式修正连接权值和阈值

$$w_{jk}(t+1) = w_{jk}(t) + \Delta w_{jk}(t)$$

对阈值，可按照连接权值的学习方式进行，只是要把阈值设想为神经元的连接权值，并假定其输入信号总为单位值 1 即可。

反复执行上述修正过程，直到满足期望的输出模式为止。

5）返回第 2）步，对训练模式集中的每一个训练模式重复第 2）到第 3）步，直到训练模式集中的每一个训练模式都满足期望输出为止。

7.2.3　MATLAB 神经网络工具箱

MATLAB 神经网络工具箱，提供了诸如：生成新网络函数、训练函数、性能函数、学习函数、预处理与后处理函数、传递函数等。

表 7-2　神经网络工具箱函数

函数名称		功　能
生成新网络函数	newcf	生成一个前向层叠 BP 网络
	newelm	生成一个 Elman BP 网络
	newff	生成一个前馈 BP 网络
	newfftd	生成前馈输入延时 BP 网络
	newhop	生成一个 Hopfield 回归网络
	newrb	设计一个径向基网络
训练函数	trainb	权重偏执学习规则成批训练
	trainbfg	BFGS 类牛顿回传
	trainbr	贝叶斯规范化
	traingdm	带动量回传的梯度递减
	trainlm	Levenberg-Marquardt 算法

（续）

函数名称		功　能
性能函数	mae	平均绝对误差性能函数
	mse	均方差性能函数
	sse	误差平方和性能函数
学习函数	learncon	公平偏执学习函数
	learngd	梯度下降权重学习函数
	learngdm	梯度下降动量权重学习函数
	learnsom	自组织映射权重学习函数
预处理与后处理函数	premnmx	规范化数据到 [−1，1]
	prestd	标准化数据（均值=0，方差=1）
	prepca	对输入数据进行主成分分析
	postmnmx	premnmx 的反函数
传递函数	logsic	对数 S 型传递函数
	poslin	正线性传递分派函数
	purelin	线性传递函数
	tansig	双曲正切 S 型传递函数

以下重点介绍几个函数的用法，其余的请读者参考 MATLAB 的在线帮助。

1）创建 BP 网络命令为 newff，其调用格式为：

```
net= newff(PR,[S1,S2,… ,SN],{TF1,TF2,… ,TFN},BTF,BLF,PF)
```

其中，PR 表示由每个输入向量的最大最小值构成的 $Rx2$ 矩阵；Si 表示第 i 层网络的神经元个数；TFi 表示第 i 层网络的传递函数，默认为 tansig，可选用的传递函数有 tansig，logsig 或 purelin；BTF 表示字符串变量，为网络的训练函数名，可在如下函数中选择：traingd、traingdm、traingdx、trainbfg、trainlm 等，默认为 trainlm；BLF 表示字符串变量，为网络的学习函数名，默认为 learngdm；BF 表示字符串变量，为网络的性能函数，默认为均方差 mse。

2）神经网络进行初始化命令为 int，其调用格式为：

```
NET= int(net)
```

其中，NET 返回参数，表示已经初始化后的神经网络；net 表示待初始化的神经网络。NET 为 net 经过一定的初始化修正而成。修正后，前者的权值和阈值都发生了改变。

3）神经网络训练命令为 train，其调用格式为：

```
[net,tr,Y,E,Pf,Af]= train(NET,p,t,Pi,Ai,)
```

其中，NET 为由 newff 产生的要训练的网络；p 和 t 分别为输入输出矩阵；Pi 为初始的输入延迟，默认为 0；Ai 为初始的层次延迟，默认为 0；net 为修正后的网络，tr 为训练的记录（训练步数 epoch 和性能 perf）；Y 函数返回值，神经网络输出信号；E 函数返回值，神经网络误差；Pf 最终输入延迟；Af 最终层延迟。

train 根据在 newff 函数中确定的训练函数来训练，不同的训练函数对应不同的训练算法。

4）均方误差性能函数 mse，其调用格式为：

```
Perf= mse(e,net,pp)
```

其中，e 为误差向量矩阵（或向量）；net 为待评定的神经网络；pp 为性能参数，可忽略。

7.2.4　BP 神经网络应用实例

1. 基于 MATLAB 的 BP 神经网络判别

基于 MATLAB 模式识别的基本步骤如下：

1）原始数据预处理，使用 premnmx（p）将数据 p 规范化到［－1，1］区间，或使用 prestd（p）将数据 p 标准化为均值为 0 方差为 1 的数据。

2）建立初始网络。

3）利用数据对网络进行训练。注意要正确地选择输入层、隐含层以及输出层函数。

4）对判别对象进行仿真识别

```
[Y,Pf,Af,E,perf]= sim(net,P,Pi,Ai,T)
```

net：使用的网络；

P：输入矩阵（判别对象）；

Pi：初始输入延迟条件，仅当输入有延迟时使用，默认为 0；

Ai：网络层初始延迟条件；

T：网络标靶；

Y：网络输出；

Pf：输入向量最终延迟条件；

Af：网络层最终延迟条件；

E：网络误差；

perf：网络的性能（每一次训练的误差）。

例 7.2.1　已知 9 个湖泊的水质观测值（见表 7-3）和湖泊水质评价标准（见表 7-4），利用 BP 神经网络进行判别。

表 7-3　全国 9 个主要湖泊评价参数的实测数据

指　标	总磷（mg/L）	总氮（mg/L）	耗氧量（mg/L）	生　物　量	透明度（m）
青海湖	20	220	1.40	14.6	4.50
太湖	20	900	2.83	100.0	0.5
呼伦湖	80	130	8.29	11.6	0.5
洪泽湖	100	460	5.5	11.5	0.3
巢湖	30	1 670	6.26	25.3	0.25
滇池	20	230	10.13	189.2	0.5
武汉东湖	105	2 000	10.7	1 913.7	0.4
杭州西湖	130	760	10.3	6 920	0.35
洱海	34	490	2.11	22.3	3.3

数据来源：韩涛，等．基于 MATLAB 的神经网络在湖泊富营养化评价中的应用．水资源保护，Vol.21，2005（1）.

表 7-4　湖泊水质评价标准

评价参数	总磷（mg/L）	总氮（mg/L）	耗氧量（mg/L）	生　物　量	透明度（m）
极贫营养	＜1	＜20	＜0.09	＜4	＞37.0
贫营养	4	60	0.36	15	12.0
中营养	23	310	1.80	50	2.4
富营养	110	1200	7.10	100	0.55
极富营养	＞660	＞4600	＞27.10	＞1 000	＜0.17

解：我们将极贫营养至极富营养 5 个等级的标靶向量（即期望输出向量）分别设置为：$(1, 0, 0, 0, 0)^T$，$(0, 1, 0, 0, 0)^T$，$(0, 0, 1, 0, 0)^T$，$(0, 0, 0, 1, 0)^T$，$(0, 0, 0, 0, 1)^T$。程序如下：

```
% 输入数据
A1= [20,220,1.40,14.6,4.50;20,900,2.83,100.0,0.5;80,130,8.29,11.6,0.5;100,460,5.5,11.5,0.3;
30,1670,6.26,25.3,0.25;20,230,10.13,189.2,0.5;105,2000,10.7,1913.7,0.4;130,760,10.3,6920,0.
35; 34,490,2.11,22.3,3.3];
p1= [1,20,0.09,4,37;4,60,0.36,15,12;23,310,1.8,50,2.4;110,1200,7.10,100,0.55; 660,4600,27.1,
1000,0.17]';
% 初始化数据
[p,minp1,maxp1]= premnmx(p1);
[A,minA1,maxA1]= premnmx(A1');                    % 将原始数据变换到[- 1,1]
net= newff(minmax(p),[8,5],{'tansig','logsig'},'traincgb','learngdm','mse');
% 建立网络
% 设置参数
net= init(net);                                   % 初始化网络
net.trainParam.epochs= 500;                       % 最大训练步数
net.trainParam.goal= 0.01;                        % 训练目标误差
net.trainParam.show= 10;                          % 每多少轮显示一次
net.trainParam.lr= 0.05;                          % 学习速度
net.trainParam.grad= 1.0e- 005;                   % 训练中最小允许梯度值
net1= train(net,p,eye(5));                        % 网络训练
r= sim(net1,A)                                    % 模拟仿真
```

各湖泊期望输出结果为：

青海湖	太 湖	呼伦湖	洪泽湖	巢湖	滇池	武汉东湖	杭州西湖	洱海
0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.85
0.09	0.02	0.02	0.00	0.00	0.02	0.00	0.01	0.18
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.99	0.99	0.99	0.87	0.98	0.04	0.02	0.00
0.00	0.00	0.00	0.00	0.12	0.00	1.00	1.00	0.00

由此可知：武汉东湖　杭州西湖属于极富营养，太湖、呼伦湖、洪泽湖、巢湖、滇池均为富营养。

2. 基于 MATLAB 的 BP 神经网络预测

利用神经网络进行预测的基本思路如下：

1）对于时间序列 $X_1, X_2, \cdots, X_n$，利用递推方法构造输入与输出，例如

输入：$X_1, X_2, X_3, \cdots, X_m$　⇒ 输出 $X_2, X_3, X_4, \cdots, X_{m+1}$；

输入：$X_2, X_3, X_4, \cdots, X_{m+1}$ ⇒ 输出 $X_3, X_4, X_5, \cdots, X_{m+2}$；

…

输入：$X_{n-m+1}, X_{n-m+2}, \cdots, X_n$ ⇒ 输出 $X_{n-m+2}, X_{n-m+3}, \cdots, X_{n+1}$

2）选择合适的网络进行训练，从而进行预测。

上述思路构造的网络适用于每次预测一个，类似地可以构造每次预测 k 个的递推公式。

神经网络预测流程图，如图 7-17 所示：

基于 MATLAB 软件神经网络预测步骤：

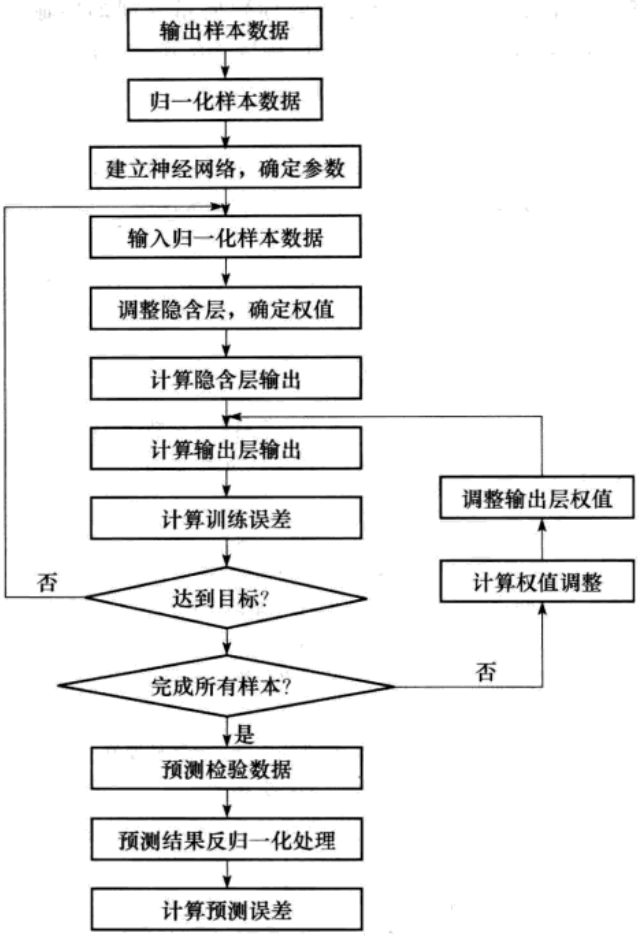


图 7-17　神经网络预测流程图

1）原始数据预处理

```
[pn1,minpn,maxpn]= premnmx(pn);
[Tn1,minTn,maxTn]= premnmx(Tn);
```

其中，*pn* 是原始数据，*Tn* 是原始数据的期望输出（即标靶）。

2）初始化网络

```
net= newff(minmax(pn1),[16,1],{'tansig','purelin'},'trainlm','learngdm');
```

3）设置网络参数

```
net.trainParam.epochs= 2500;      % 最大步长
net.trainParam.goal= 0.001;       % 精度
net.trainParam.show= 10;          % 显示两次间隔间的次数
net.trainParam.lr= 0.8;           % 学习率
net.trainParam.mc= 0.6;           % 动量因子
```

4）对网络进行训练

```
net= train(net,pn1,Tn1);
```

5）原始数据用网络进行仿真

```
rn1= sim(net,pn1);
```

6）将仿真后的数据还原

```
y= postmnmx(rn1,minTn,maxTn)
```

例 7.2.2　根据 1949—1990 年的受灾数据（见表 7-5）对受灾面积进行预测。

表 7-5　1949—1990 年受灾数据

年　份	受灾面积	受灾人口	直接经济损失	年　份	受灾面积	受灾人口	直接经济损失
1949	928.2	2 006	190 300	1970	313	305	17 424.71
1950	656	1 928	12 028.87	1971	399	618	15 312.09
1951	417	601	12 614.71	1972	408	1 608	21 804
1952	279.4	1 059	23 339.56	1973	624	1 746	14 378.77
1953	741	812	10 897.38	1974	640	1 988	35 974.6
1954	1613	3 937	209 300	1975	682	1 208	1 000 000
1955	525	407	13 061.56	1976	420	2 589	26 163.63
1956	1 438	2 576	326 801.7	1977	910	1 872	60 604.77
1957	808.27	870	45 708.41	1978	285	2 130	26 155.93
1958	428	1 132	14 692	1979	676	2 191	54 798.1
1959	481	845	25 746	1980	915	4 106	90 339.39
1960	1 016	682	58 179.59	1981	862	4 560	335 319.3
1961	887	1 867	26 172.85	1982	836	4 499	120 239.5
1962	981	1 501	53 865.8	1983	1 216	5 294	221 760.3
1963	1 407	2 757	629 755.2	1984	1 069	nan	1 530
1964	1 493	1 561	31 458.73	1985	1 419.73	1 294	470 282
1965	559	683	2 3751.14	1986	915.53	321	703 600
1966	251	1 079	68 286.03	1987	868.6	2 105	246 253.3
1967	170.89	575	14 286.03	1988	1 194.93	3 522	803 387.8
1968	224.34	372	8 232.32	1989	1 132.8	nan	233 000
1969	463.18	1 252	23 293.55	1990	1 180.4	7 611	1 591 968

数据来源：李柏年．洪灾损失的回归模型［J］．昆明理工大学学报，2006，31(2).

解：将 1949—1988 年的数据作为训练样本，1989 年、1990 年数据作为检验。

首先将受灾面积数据输入，并记为 a，然后输入以下程序：

```
p= [a(1:32);a(2:33);a(3:34);a(4:35);a(5:36);a(6:37);a(7:38);a(8:39)];
T= [a(2:33);a(3:34);a(4:35);a(5:36);a(6:37);a(7:38);a(8:39);a(9:40)]';
[p1,minp,maxp]= premnmx(p);
[T1,minT,maxT]= premnmx(T');
net= newff(minmax(p1),[17,8],{ 'logsig','purelin'},'traincgp','learnwh');
net.trainParam.epochs= 2500;
net.trainParam.goal= 0.001;
net.trainParam.show= 10;
net.trainParam.lr= 0.8;
net.trainParam.mc= 0.6;
net= train(net,p1,T1);
r1= sim(net,p1);
yu= postmnmx(r1,minT,maxT);
y1= [a(1),yu(1,1:32),yu(2,32),yu(3,32),yu(4,32),yu(5,32),yu(6,32),yu(7,32),yu(8,32)];
e1= (y1- a(1:40));
f1 = mae(e1);
```

```
f2 = mse(e1);
f3 = sse(e1);
F1=[f1,f2,f3]
pt1=[a(2:33);a(3:34);a(4:35);a(5:36);a(6:37);a(7:38);a(8:39);a(9:40)];
[py1,minpt1,maxpt1]= premnmx(pt1);
ry1= sim(net,py1);
yc1= postmnmx(ry1,minpt1,maxpt1);
yu1= yc1(8,32)
pt2=[a(3:34);a(4:35);a(5:36);a(6:37);a(7:38);a(8:39);a(9:40);[a(10:40),yu1]];
[py2,minpt2,maxpt2]= premnmx(pt2);
ry2= sim(net,py2);
yc2= postmnmx(ry2,minpt2,maxpt2);
yu2= yc2(8,32)
```

模型预测结果误差分析见表 7-6。

表 7-6　预测结果误差分析表

年　份	实际数据	预测结果	相对误差
1989	1 132.8	1 197.7	0.057 3
1990	1 180.4	1 068.4	−0.094 9

由于检验结果较好，可以利用全部数据进行预测，此时要将原程序稍加改动，预测 1991 年、1992 年两年的受灾面积。

受灾面积实际数据与预测数据图形，如图 7-18 所示。

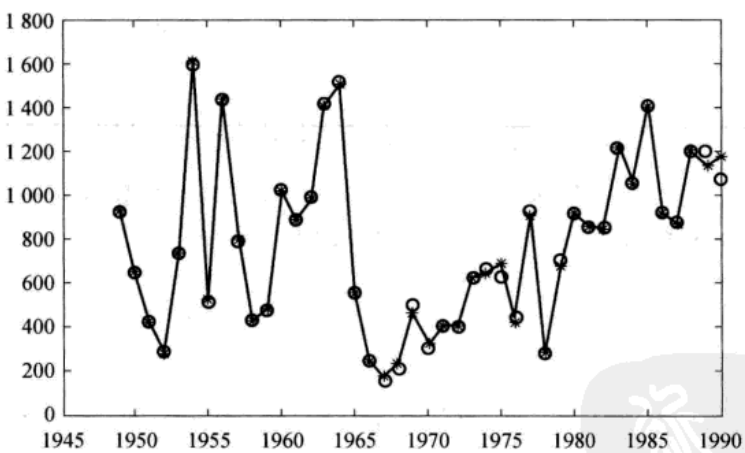


图 7-18　受灾面积实际数据（*）与预测数据（o）图形

习　题　7

1. 应用蒙特卡罗方法计算定积分 $\int_0^1 e^{x^2}\,dx$。

2. 设 $U_1, U_2, \cdots$ 独立同分布且都在（0，2π）上均匀分布

$$X_t = b\cos(at + U_t),\quad t \in Z$$

计算 $E(X_t)$，$D(X_t)$；任意给定 a、b 的值，模拟生成序列 $\{X_t\}$ 的 300 个样本，求出样本均值与标

准差，并与真实值 $E(X_t)$，$D(X_t)$ 作比较。

3. 设 ARMA(2，1) 模型为 $x_t=0.2x_{t-1}-0.5x_{t-2}+\varepsilon_t-0.5\varepsilon_{t-1}$，$t\in Z$，模拟 $\{x_t\}$ 的 500 个观测值，求出样本的自协方差函数，由模拟样本估计谱密度。将估计谱密度与真实谱密度函数作比较。
4. 应用 BP 神经网络，对习题 2 的第 4 题建立预测模型，预测 2005 年的各项指标的值。

实验 6　数值模拟

实验目的

1. 熟练掌握利用 MATLAB 软件生成各种分布的随机数的方法。
2. 掌握蒙特卡罗方法的应用。
3. 掌握基于 MATLAB 的 BP 神经网络的创建与模拟计算。

实验数据与内容

汇率是将一个国家的货币折算成另一个国家货币时使用的折算比率，也可以说是货币的相对价格。它在本质上反映的是不同国家货币之间的价值对比关系。以本币来表示单位外币的价格叫直接标价（the direct quotation），简称为直接汇率；用外币来表示单位本币的价格叫间接标价（the indirect quotation），简称为间接汇率。汇率作为一个重要的经济变量，其变动对国民收入的增减、工农业的发展、国内利率、就业等各方面都有着重要的影响。尤其是在全球经济一体化趋势逐渐加强和世界各国经济之间的依赖程度不断加深的今天，汇率无疑成了维系国际间经济往来的纽带和桥梁，具有越来越重要的地位。汇率的决定及其变化也对国际贸易、国际投资、国际收支等产生重大影响。从国外汇管理局网站 http：//www. safe. gov. cn/model_safe/index. html，获取人民币对美元、欧元、日元、英镑汇率的日中间价数据，时间为 2008 年 10 月 6 日至 12 月 31 日（见表 7-7），建立神经网络模型预测 2009 年 1 月 1 日至 10 日的汇率。

表 7-7　人民币汇率日中间价（100 外币/人民币）

日　期	美　元	欧　元	日　元	英　镑
2008-12-31	683.46	965.9	7.565	987.98
2008-12-30	683.53	961.04	7.520 4	989.75
2008-12-29	683.57	965.06	7.537 4	1 004.71
2008-12-26	683.62	957.24	7.528	1 007.18
2008-12-25	683.66	956.78	7.559 3	1 008.6
2008-12-24	683.97	953.83	7.540 2	1 009.1
2008-12-23	683.89	955.57	7.581 1	1 013.8
2008-12-22	683.77	957.35	7.601 7	1 023.19
2008-12-19	683.57	975.69	7.609 2	1 031.06
2008-12-18	683.22	985.61	7.814 5	1 061.01
2008-12-17	683.53	960.6	7.680 5	1 064.94
2008-12-16	684.33	937.09	7.555	1 046.65
2008-12-15	684.42	920.89	7.507 9	1 027.9

（续）

日　期	美　元	欧　元	日　元	英　镑
2008-12-12	684.51	911.8	7.477 7	1 027.72
2008-12-11	684.71	891.56	7.392 3	1 014.26
2008-12-10	684.75	885.59	7.395 1	1 011.14
2008-12-09	684.79	884.68	7.365 7	1 018.97
2008-12-08	685.09	874.28	7.391 2	1 008.52
2008-12-05	684.82	872.87	7.402 3	1 003.57
2008-12-04	685.02	869.15	7.349 6	1 010.54
2008-12-03	685.02	869.02	7.370 6	1 019.41
2008-12-02	685.27	864.71	7.310 3	1 022.18
2008-12-01	685.05	867.2	7.186 5	1 050.66
2008-11-28	683.49	881.19	7.158 8	1 051.41
2008-11-27	682.92	879.81	7.160 7	1 047.7
2008-11-26	682.72	888.29	7.184 3	1 050.98
2008-11-25	682.84	878.75	7.065 4	1 032.28
2008-11-24	683.04	861.21	7.171 4	1 014.86
2008-11-21	683.17	849.9	7.251 6	1 007.78
2008-11-20	683.07	854.69	7.127 6	1 023.92
2008-11-19	682.93	862.34	7.074 8	1 019.82
2008-11-18	682.8	861.35	7.077 8	1 021.67
2008-11-17	683.03	856.86	7.079 1	1 003.85
2008-11-14	682.89	871.37	7.032 5	1 013.37
2008-11-13	682.95	850.82	7.142 7	1 016.78
2008-11-12	682.91	856.37	6.999 9	1 052.57
2008-11-11	682.65	866.83	7.000 8	1 064.15
2008-11-10	682.52	879.32	6.881 6	1 081.32
2008-11-07	682.77	864.97	7.033 1	1 061.88
2008-11-06	682.52	880.96	6.954 6	1 083.26
2008-11-05	682.4	883.78	6.842 1	1 087.03
2008-11-04	682.61	859.68	6.889 5	1 076.1
2008-11-03	682.88	873.37	6.935 3	1 100.7
2008-10-31	682.58	872.54	6.946 3	1 111.31
2008-10-30	682.7	892.97	6.931	1 124.48
2008-10-29	683.18	871.02	6.980 5	1 095.58
2008-10-28	683.69	848.9	7.318 8	1 058.01
2008-10-27	683.6	861.2	7.261 1	1 082.14
2008-10-24	683.57	879.45	7.028 6	1 102.22
2008-10-23	683.68	875.01	7.004 9	1 108.42

（续）

日　期	美　元	欧　元	日　元	英　镑
2008-10-22	683.39	890.15	6.806	1 133.71
2008-10-21	683.09	910.15	6.701 2	1 173.28
2008-10-20	682.97	919.62	6.709 6	1 184.1
2008-10-17	683.11	920.56	6.723 2	1 184.27
2008-10-16	682.95	920.1	6.828 5	1 179.76
2008-10-15	682.72	927.44	6.717	1 188.68
2008-10-14	682.78	930.94	6.641 5	1 192.03
2008-10-13	682.87	926.83	6.778 9	1 167.95
2008-10-10	683.27	926.34	6.901	1 163.27
2008-10-09	683.1	928.23	6.843	1 175.21
2008-10-08	683.19	928.42	6.725 6	1 193.94
2008-10-07	683.45	923.82	6.725 9	1 192.38
2008-10-06	683.21	933.03	6.547 3	1 206.48

数据来源：http：//www.safe.gov.cn/model_safe/index.html.

附录

瑞士银行纸币（Swiss Bank Notes）

附表 1　瑞士银行 1000 法郎真假纸币的数据

长　度	左侧高度	右侧高度	图廓下边距	图廓上边距	对角线长度
214.8	131.0	131.1	9.0	9.7	141.0
214.6	129.7	129.7	8.1	9.5	141.7
214.8	129.7	129.7	8.7	9.6	142.2
214.8	129.7	129.6	7.5	10.4	142.0
215.0	129.6	129.7	10.4	7.7	141.8
215.7	130.8	130.5	9.0	10.1	141.4
215.5	129.5	129.7	7.9	9.6	141.6
214.5	129.6	129.2	7.2	10.7	141.7
214.9	129.4	129.7	8.2	11.0	141.9
215.2	130.4	130.3	9.2	10.0	140.7
215.3	130.4	130.3	7.9	11.7	141.8
215.1	129.5	129.6	7.7	10.5	142.2
215.2	130.8	129.6	7.9	10.8	141.4
214.7	129.7	129.7	7.7	10.9	141.7
215.1	129.9	129.7	7.7	10.8	141.8
214.5	129.8	129.8	9.3	8.5	141.6
214.6	129.9	130.1	8.2	9.8	141.7
215.0	129.9	129.7	9.0	9.0	141.9
215.2	129.6	129.6	7.4	11.5	141.5
214.7	130.2	129.9	8.6	10.0	141.9
215.0	129.9	129.3	8.4	10.0	141.4
215.6	130.5	130.0	8.1	10.3	141.6
215.3	130.6	130.0	8.4	10.8	141.5
215.7	130.2	130.0	8.7	10.0	141.6
215.1	129.7	129.9	7.4	10.8	141.1
215.3	130.4	130.4	8.0	11.0	142.3
215.5	130.2	130.1	8.9	9.8	142.4
215.1	130.3	130.3	9.8	9.5	141.9
215.1	130.0	130.0	7.4	10.5	141.8
214.8	129.7	129.3	8.3	9.0	142.0
215.2	130.1	129.8	7.9	10.7	141.8

（续）

长　度	左侧高度	右侧高度	图廓下边距	图廓上边距	对角线长度
214.8	129.7	129.7	8.6	9.1	142.3
215.0	130.0	129.6	7.7	10.5	140.7
215.6	130.4	130.1	8.4	10.3	141.0
215.9	130.4	130.0	8.9	10.6	141.4
214.6	130.2	130.2	9.4	9.7	141.8
215.5	130.3	130.0	8.4	9.7	141.8
215.3	129.9	129.4	7.9	10.0	142.0
215.3	130.3	130.1	8.5	9.3	142.1
213.9	130.3	129.0	8.1	9.7	141.3
214.4	129.8	129.2	8.9	9.4	142.3
214.8	130.1	129.6	8.8	9.9	140.9
214.9	129.6	129.4	9.3	9.0	141.7
214.9	130.4	129.7	9.0	9.8	140.9
214.8	129.4	129.1	8.2	10.2	141.0
214.3	129.5	129.4	8.3	10.2	141.8
214.8	129.9	129.7	8.3	10.2	141.5
214.8	129.9	129.7	7.3	10.9	142.0
214.6	129.7	129.8	7.9	10.3	141.1
214.5	129.0	129.6	7.8	9.8	142.0
214.6	129.8	129.4	7.2	10.0	141.3
215.3	130.6	130.0	9.5	9.7	141.1
214.5	130.1	130.0	7.8	10.9	140.9
215.4	130.2	130.2	7.6	10.9	141.6
214.5	129.4	129.5	7.9	10.0	141.4
215.2	129.7	129.4	9.2	9.4	142.0
215.7	130.0	129.4	9.2	10.4	141.2
215.0	129.6	129.4	8.8	9.0	141.1
215.1	130.1	129.9	7.9	11.0	141.3
215.1	130.0	129.8	8.2	10.3	141.4
215.1	129.6	129.3	8.3	9.9	141.6
215.3	129.7	129.4	7.5	10.5	141.5
215.4	129.8	129.4	8.0	10.6	141.5
214.5	130.0	129.5	8.0	10.8	141.4
215.0	130.0	129.8	8.6	10.6	141.5
215.2	130.6	130.0	8.8	10.6	140.8
214.6	129.5	129.2	7.7	10.3	141.3
214.8	129.7	129.3	9.1	9.5	141.5
215.1	129.6	129.8	8.6	9.8	141.8
214.9	130.2	130.2	8.0	11.2	139.6
213.8	129.8	129.5	8.4	11.1	140.9

（续）

长　度	左侧高度	右侧高度	图廓下边距	图廓上边距	对角线长度
215.2	129.9	129.5	8.2	10.3	141.4
215.0	129.6	130.2	8.7	10.0	141.2
214.4	129.9	129.6	7.5	10.5	141.8
215.2	129.9	129.7	7.2	10.6	142.1
214.1	129.6	129.3	7.6	10.7	141.7
214.9	129.9	130.1	8.8	10.0	141.2
214.6	129.8	129.4	7.4	10.6	141.0
215.2	130.5	129.8	7.9	10.9	140.9
214.6	129.9	129.4	7.9	10.0	141.8
215.1	129.7	129.7	8.6	10.3	140.6
214.9	129.8	129.6	7.5	10.3	141.0
215.2	129.7	129.1	9.0	9.7	141.9
215.2	130.1	129.9	7.9	10.8	141.3
215.4	130.7	130.2	9.0	11.1	141.2
215.1	129.9	129.6	8.9	10.2	141.5
215.2	129.9	129.7	8.7	9.5	141.6
215.0	129.6	129.2	8.4	10.2	142.1
214.9	130.3	129.9	7.4	11.2	141.5
215.0	129.9	129.7	8.0	10.5	142.0
214.7	129.7	129.3	8.6	9.6	141.6
215.4	130.0	129.9	8.5	9.7	141.4
214.9	129.4	129.5	8.2	9.9	141.5
214.5	129.5	129.3	7.4	10.7	141.5
214.7	129.6	129.5	8.3	10.0	142.0
215.6	129.9	129.9	9.0	9.5	141.7
215.0	130.4	130.3	9.1	10.2	141.1
214.4	129.7	129.5	8.0	10.3	141.2
215.1	130.0	129.8	9.1	10.2	141.5
214.7	130.0	129.4	7.8	10.0	141.2
214.4	130.1	130.3	9.7	11.7	139.8
214.9	130.5	130.2	11.0	11.5	139.5
214.9	130.3	130.1	8.7	11.7	140.2
215.0	130.4	130.6	9.9	10.9	140.3
214.7	130.2	130.3	11.8	10.9	139.7
215.0	130.2	130.2	10.6	10.7	139.9
215.3	130.3	130.1	9.3	12.1	140.2
214.8	130.1	130.4	9.8	11.5	139.9
215.0	130.2	129.9	10.0	11.9	139.4
215.2	130.6	130.8	10.4	11.2	140.3
215.2	130.4	130.3	8.0	11.5	139.2

（续）

长　度	左侧高度	右侧高度	图廓下边距	图廓上边距	对角线长度
215.1	130.5	130.3	10.6	11.5	140.1
215.4	130.7	131.1	9.7	11.8	140.6
214.9	130.4	129.9	11.4	11.0	139.9
215.1	130.3	130.0	10.6	10.8	139.7
215.5	130.4	130.0	8.2	11.2	139.2
214.7	130.6	130.1	11.8	10.5	139.8
214.7	130.4	130.1	12.1	10.4	139.9
214.8	130.5	130.2	11.0	11.0	140.0
214.4	130.2	129.9	10.1	12.0	139.2
214.8	130.3	130.4	10.1	12.1	139.6
215.1	130.6	130.3	12.3	10.2	139.6
215.3	130.8	131.1	11.6	10.6	140.2
215.1	130.7	130.4	10.5	11.2	139.7
214.7	130.5	130.5	9.9	10.3	140.1
214.9	130.0	130.3	10.2	11.4	139.6
215.0	130.4	130.4	9.4	11.6	140.2
215.5	130.7	130.3	10.2	11.8	140.0
215.1	130.2	130.2	10.1	11.3	140.3
214.5	130.2	130.6	9.8	12.1	139.9
214.3	130.2	130.0	10.7	10.5	139.8
214.5	130.2	129.8	12.3	11.2	139.2
214.9	130.5	130.2	10.6	11.5	139.9
214.6	130.2	130.4	10.5	11.8	139.7
214.2	130.0	130.2	11.0	11.2	139.5
214.8	130.1	130.1	11.9	11.1	139.5
214.6	129.8	130.2	10.7	11.1	139.4
214.9	130.7	130.3	9.3	11.2	138.3
214.6	130.4	130.4	11.3	10.8	139.8
214.5	130.5	130.2	11.8	10.2	139.6
214.8	130.2	130.3	10.0	11.9	139.3
214.7	130.0	129.4	10.2	11.0	139.2
214.6	130.2	130.4	11.2	10.7	139.9
215.0	130.5	130.4	10.6	11.1	139.9
214.5	129.8	129.8	11.4	10.0	139.3
214.9	130.6	130.4	11.9	10.5	139.8
215.0	130.5	130.4	11.4	10.7	139.9
215.3	130.6	130.3	9.3	11.3	138.1
214.7	130.2	130.1	10.7	11.0	139.4
214.9	129.9	130.0	9.9	12.3	139.4
214.9	130.3	129.9	11.9	10.6	139.8

（续）

长　度	左侧高度	右侧高度	图廓下边距	图廓上边距	对角线长度
214.6	129.9	129.7	11.9	10.1	139.0
214.6	129.7	129.3	10.4	11.0	139.3
214.5	130.1	130.1	12.1	10.3	139.4
214.5	130.3	130.0	11.0	11.5	139.5
215.1	130.0	130.3	11.6	10.5	139.7
214.2	129.7	129.6	10.3	11.4	139.5
214.4	130.1	130.0	11.3	10.7	139.2
214.8	130.4	130.6	12.5	10.0	139.3
214.6	130.6	130.1	8.1	12.1	137.9
215.6	130.1	129.7	7.4	12.2	138.4
214.9	130.5	130.1	9.9	10.2	138.1
214.6	130.1	130.0	11.5	10.6	139.5
214.7	130.1	130.2	11.6	10.9	139.1
214.3	130.3	130.0	11.4	10.5	139.8
215.1	130.3	130.6	10.3	12.0	139.7
216.3	130.7	130.4	10.0	10.1	138.8
215.6	130.4	130.1	9.6	11.2	138.6
214.8	129.9	129.8	9.6	12.0	139.6
214.9	130.0	129.9	11.4	10.9	139.7
213.9	130.7	130.5	8.7	11.5	137.8
214.2	130.6	130.4	12.0	10.2	139.6
214.8	130.5	130.3	11.8	10.5	139.4
214.8	129.6	130.0	10.4	11.6	139.2
214.8	130.1	130.0	11.4	10.5	139.6
214.9	130.4	130.2	11.9	10.7	139.0
214.3	130.1	130.1	11.6	10.5	139.7
214.5	130.4	130.0	9.9	12.0	139.6
214.8	130.5	130.3	10.2	12.1	139.1
214.5	130.2	130.4	8.2	11.8	137.8
215.0	130.4	130.1	11.4	10.7	139.1
214.8	130.6	130.6	8.0	11.4	138.7
215.0	130.5	130.1	11.0	11.4	139.3
214.6	130.5	130.4	10.1	11.4	139.3
214.7	130.2	130.1	10.7	11.1	139.5
214.7	130.4	130.0	11.5	10.7	139.4
214.5	130.4	130.0	8.0	12.2	138.5
214.8	130.0	129.7	11.4	10.6	139.2
214.8	129.9	130.2	9.6	11.9	139.4
214.6	130.3	130.2	12.7	9.1	139.2
215.1	130.2	129.8	10.2	12.0	139.4

（续）

长　度	左侧高度	右侧高度	图廓下边距	图廓上边距	对角线长度
215.4	130.5	130.6	8.8	11.0	138.6
214.7	130.3	130.2	10.8	11.1	139.2
215.0	130.5	130.3	9.6	11.0	138.5
214.9	130.3	130.5	11.6	10.6	139.8
215.0	130.4	130.3	9.9	12.1	139.6
215.1	130.3	129.9	10.3	11.5	139.7
214.8	130.3	130.4	10.6	11.1	140.0
214.7	130.7	130.8	11.2	11.2	139.4
214.3	129.9	129.9	10.2	11.5	139.6

数据提供者：弗洛里（Flury）和里德威尔（Riedwyl）（1998）。

参考文献

[1] Richard A Johnson, Dean, W Wichern. 应用多元统计分析 [M]. 英文版. 北京：中国统计出版社，2002.

[2] 何晓群. 多元统计分析 [M]. 北京：中国人民大学出版社，2004.

[3] 张尧庭，方开泰. 多元统计分析引论 [M]. 北京：科学出版社，1999.

[4] 范金城，梅长林. 数据分析 [M]. 北京：科学出版社，2002.

[5] 高惠璇. 应用多元统计分析 [M]. 北京：北京大学出版社，2005.

[6] James M Lattin, J Douglas Carroll, Paul E Green. 多元数据分析 [M]. 英文版. 北京：机械工业出版社，2003.

[7] 吴礼斌，李柏年. 数学实验与建模 [M]. 北京：国防工业出版社，2007.

[8] 茆诗松. 贝叶斯统计 [M]. 北京：中国统计出版社，1999.

[9] 张立军，任英华. 多元统计分析实验 [M]. 北京：中国统计出版社，2009.

[10] 向东进. 实用多元统计分析 [M]. 武汉：中国地质大学出版社，2005.

[11] 王岩，隋思涟，王爱青. 数理统计与 MATLAB 工程数据分析 [M]. 北京：清华大学出版社，2006.

[12] 邓留保，李柏年，杨桂元. Matlab 与金融模型分析 [M]. 合肥：合肥工业大学出版社，2007.

[13] 梅长林，范金城. 数据分析方法 [M]. 北京：高等教育出版社，2006.

[14] 范九伦. 模糊聚类新算法与聚类有效性问题研究 [D]. 西安：西安电子科技大学，1998.

[15] 李柏年. 模糊数学及其应用 [M]. 合肥：合肥工业大学出版社，2007.

[16] D W Kim, K Y Lee, D Lee, K H Lee. A kernel-based subtractive clustering method [J]. Pattern Recognition Letters, 2005, 26: 879-891.

[17] W N Wang, Y J Zhang. On fuzzy cluster validity indices [J]. Fuzzy Sets and Systems, 2007, 158: 2095-2117.

[18] S L Chiu. Extracting fuzzy rules for pattern classification by cluster estimation [C]. In: The 6th International. Fuzzy Systems Association World Congress, 1995: 1-4.

[19] H Sarimveis, A Alexandridis, G Bafas. A fast training algorithm for RBF networks based on subtractive clustering. Neurocomputing, 2003, 51: 501-505.

[20] 李柏年. 经济数据处理与优化模型分析实验教程 [M]. 天津：天津大学出版社，2009.

[21] 李琼，周建中. 改进主成分分析法在洪灾损失评估中的应用 [J]. 水电能源科学，2010，28 (3)：39-42.

[22] 王秀峰，卢桂章. 系统建模与辨识 [M]. 北京：电子工业出版社，2004.

[23] 何晓群，刘文卿. 应用回归分析 [M]. 北京：中国人民大学出版社，2001.

[24] 飞思科技产品研发中心. 神经网络理论与 MATLAB 7 实现 [M]. 北京：电子工业出版社，2005.

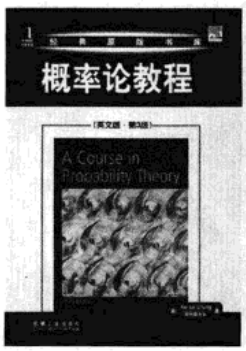
[25]《现代应用数学手册》编委会. 现代应用数学手册：概率统计与随机过程卷 [M]. 北京：清华大学出版社，2002.

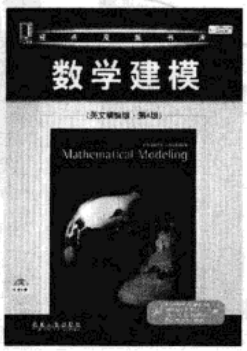
[26] 高祥宝，董寒青. 数据分析与 SPSS 应用 [M]. 北京：清华大学出版社，2007.

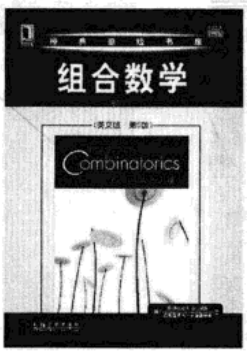
[27] 郑阿奇，曹弋. MATLAB 实用教程 [M]. 2 版. 北京：电子工业出版社，2007.

[28] Statistics Toolbox User's Guide. http: //www. mathworks. com.

华章数学经典原版

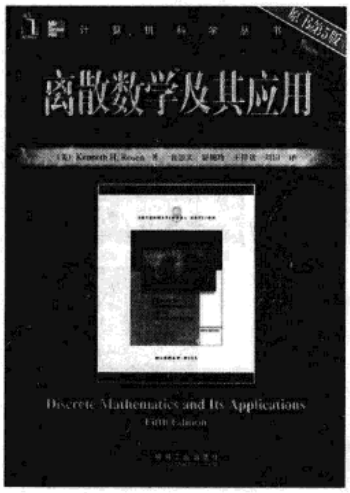




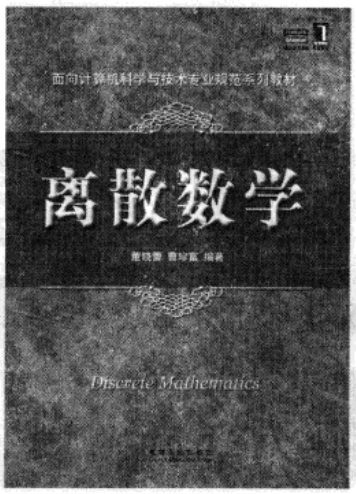


书名	书号	定价	出版年	作者
离散数学及其应用 （英文精编版. 第6版）	978-7-111-31329	55	2010	（美）Kenneth H. Rosen
初等数论及其应用（英文版 第6版）	978-7-111-31792	89	2010	（美）Kenneth H. Rosen
统计模型：理论和实践（英文版 第2版）	978-7-111-31797	38	2010	（美）David A.Freedman
概率论教程 （英文版 第3版）	978-7-111-30289	49	2010	（美）Kai LaiChung
数学建模（英文精编版 第4版）	978-7-111-28249	65	2009	（美）FrankR. Giordano
组合数学 （英文版 第5版）	978-7-111-26525	49	2009	（美）Richard A. Brualdi
复变函数及应用 （英文版 第8版）	978-7-111-25363	65	2009	（美）James Ward Brown
算法概论 （注释版）	978-7-111-25361	55	2009	（美）Sanjoy Dasgupts 钱枫注译
数学建模 方法与分析（英文版. 第3版）	978-7-111-25364	49	2008	（美）MarkM. Meerschaert
离散数学及其应用 （英文版. 第6版）	978-7-111-23935	89	2008	（美）Kenneth H. Rosen
线性代数（英文版. 第7版）	978-7-111-21198	58	2007	（美）Steven J. Leon
离散数学（英文版. 第5版）	978-7-111-20167	75	2006	（美）Lawrence E. Spence Charles Vanden Eynd
应用逻辑（英文版. 第2版）	978-7-111-19772	49	2006	（美）Anil Nerode Richard A. Shore
数论概论（英文版. 第3版）	978-7-111-19611	52	2006	（美）Joseph H. Silverman
高等微积分（英文版. 第2版）	978-7-111-19349	76	2006	（美）Patrick M. Fitzpatrick
实分析和概率论（英文版. 第2版）	978-7-111-19348	69	2006	（美） R. M. Dudley

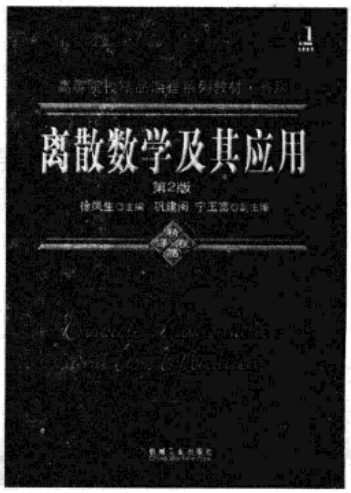
好书推荐



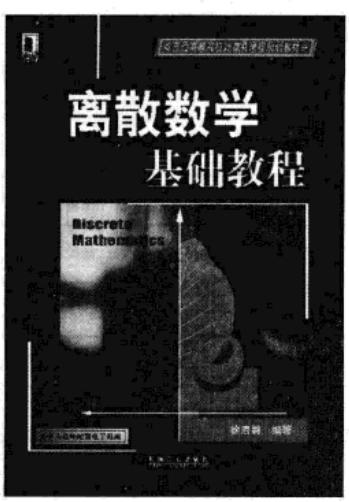
作者：Kenneth H. Rosen（AT&T实验室）
译者：袁崇义 屈婉玲 王捍贫 刘田（北京大学）
本科教学版6/e 2010 预计出版
英文版6/e 2008
ISBN：7-111-23935-2 定价：89.00
ISBN：7-111-20326-1 定价：79.00



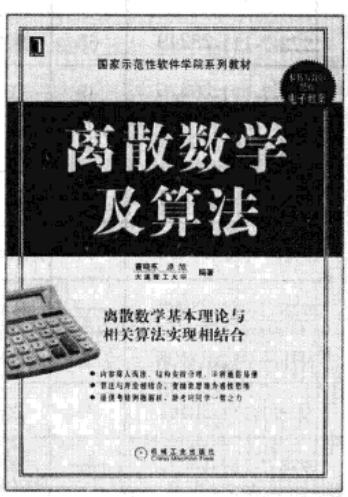
作者：董晓蕾 曹珍富
ISBN：7-111-23571-2
定价：35.00



作者：徐凤生 等编著
ISBN：7-111-27284-7
定价：30.00



作者：徐洁磐
ISBN：7-111-27431-5
定价：29.00



作者：曹晓东 原 旭
ISBN：7-111-21876-0
定价：28.00

作者：林成森 史九林 周建兰 徐进鸿 徐洁磐
ISBN：978-7-111-29854-0
定价：35.00

教师服务登记表

尊敬的老师：

您好！感谢您购买我们出版的_____教材。

机械工业出版社华章公司为了进一步加强与高校教师的联系与沟通，更好地为高校教师服务，特制此表，请您填妥后发回给我们，我们将定期向您寄送华章公司最新的图书出版信息！感谢合作！

个人资料（请用正楷完整填写）

<table>
<tr><td>教师姓名</td><td></td><td>□先生
□女士</td><td>出生年月</td><td></td><td>职务</td><td></td><td colspan="2">职称：□教授 □副教授
□讲师 □助教 □其他</td></tr>
<tr><td>学校</td><td colspan="2"></td><td>学院</td><td colspan="2"></td><td>系别</td><td colspan="2"></td></tr>
<tr><td rowspan="2">联系
电话</td><td colspan="3" rowspan="2">办公：
宅电：
移动：</td><td>联系地址
及邮编</td><td colspan="3"></td></tr>
<tr><td>E-mail</td><td colspan="3"></td></tr>
<tr><td>学历</td><td></td><td>毕业院校</td><td colspan="2"></td><td>国外进修及讲学经历</td><td colspan="3"></td></tr>
<tr><td>研究领域</td><td colspan="8"></td></tr>
<tr><td colspan="2">主讲课程</td><td colspan="3">现用教材名</td><td>作者及
出版社</td><td>共同授
课教师</td><td colspan="2">教材满意度</td></tr>
<tr><td colspan="2">课程：
□专 □本 □研
人数： 学期：□春□秋</td><td colspan="3"></td><td></td><td></td><td colspan="2">□满意 □一般
□不满意 □希望更换</td></tr>
<tr><td colspan="2">课程：
□专 □本 □研
人数： 学期：□春□秋</td><td colspan="3"></td><td></td><td></td><td colspan="2">□满意 □一般
□不满意 □希望更换</td></tr>
<tr><td colspan="9">样书申请</td></tr>
<tr><td colspan="3">已出版著作</td><td colspan="3"></td><td colspan="2">已出版译作</td><td></td></tr>
<tr><td colspan="5">是否愿意从事翻译/著作工作 □是 □否</td><td>方向</td><td colspan="3"></td></tr>
<tr><td>意见和建议</td><td colspan="8"></td></tr>
</table>

填妥后请选择以下任何一种方式将此表返回：（如方便请赐名片）

地 址：北京市西城区百万庄南街1号 华章公司营销中心 邮编：100037

电 话：(010) 68353079 88378995 传真：(010)68995260

E-mail:hzedu@hzbook.com markerting@hzbook.com 图书详情可登录http://www.hzbook.com网站查询