统计计算

李东风

2017 年 7 月 5 日

目录

第一章　绪论

1.1　介绍

1.1.1　统计计算的范畴

统计计算是现代统计的重要组成部分。从上个世纪三、四十年代起，数理统计的理论和方法得到跨越式的发展，统计推断理论、回归分析、试验设计、方差分析、序贯分析、时间序列分析、随机过程等理论和方法在这个时期逐渐成熟。但是，直到上个世纪八十年代，统计学作为一门学科才真正得到了广泛的普及，其应用深入到了我们的学术研究和社会生活的每一个方面，只要需要分析数据的地方就需要用到统计学。这种普及很大程度上要归功于电子信息技术的高速发展。

统计计算就是统计方法和实际计算的结合。统计计算有以下两个方面的内容：

- 把统计方法变成可靠、高效的算法，并编程实现。这是经典的统计计算要解决的问题，比如计算分布函数值、分位数函数值、计算线性回归参数估计和检验、求解最大似然估计等。
- 借助于现代计算机的强大处理能力，发展新的统计方法。这是计算技术对统计学的贡献，比如用随机模拟方法求解贝叶斯模型、Bootstrap 置信区间，等等。这个方面有时被称为 “计算统计”(computational statistics)。

第二个方面的内容包括计算密集的统计方法及相应的理论工作。**随机模拟**方法是其中一个重要内容。随机模拟的基本思想是在计算机上模拟生成一个统计问题的数据并进行大量的重复，这样相当于获得了此问题的海量的样本。如果我们的目的是评估某种建模方法，我们可以对每个样本建模，最后从所有建模结果的评估这种方法的性能。可以从理论模型产生海量模拟样本后对此模型进行理论推断，如蒙特卡洛检验。可以对观测数据进行多次重复再抽样生成许多新样本，如 Bootstrap 方法。在贝叶斯统计框架下，我们可以从先验分布抽样并

按照模型产生大量样本并结合观测数据计算其似然，从而获得参数后验分布的大量样本，以此进行贝叶斯推断。借助于随机模拟，我们可以试验各种各样的模型与方法，发现表现优秀的模型和方法后再进行深入研究。

另外，各行各业中数据收集越来越广泛，在迅速积累的海量数据中包含了许多以前无法触摸的现象和规律，对海量数据进行探索性分析，从海量数据中发现规律，已经成为统计学和信息科学的热门研究方法，通常称为机器学习、数据挖掘等。这也是统计计算第二个方面的重要内容。

现在已经有了许多专用的统计软件，比如 R、SAS 等，为我们平常遇到的许多问题提供了现成的解决办法，那么，我们为什么还需要学习统计计算呢？

我们遇到的具体应用问题常常是没有现成的方法可以套用的。即使有现成的统计软件可用，我们也需要理解这些软件的工作原理以避免错误使用；在遇到新问题时，需要能够修改原有代码或编写新代码，把计算工具结合在一起解决自己的数据分析问题，而不是修改自己的问题以适应现成的软件。

1.1.2 算法和计算机语言

算法是完成某项任务的步骤的精确的描述。比如，制作水果沙拉的一种算法为：

> 1．准备一个盘子；
> 2．取一个苹果削皮后把果肉切成小块放入盘中；
> 3．取一个香蕉去皮后切成小块放入盘中；
> 4．取一调羹沙拉酱放入盘中并与果料一起搅拌均匀。

当然，算法主要针对在电子计算机上的计算而设计。好的算法应该满足如下要求：

- 结果正确，即要求算法的最后结果是我们问题的正确解，最好能够验证结果的正确性。
- 指令可行，即指令含义明确无歧义，指令可以执行并且在现有的计算条件下算法能在允许的时间计算结束。
- 高效，尽可能少地消耗时间和内存、外存资源。

电子计算机由 CPU、内存、大容量外存、输入/输出装置等硬件构成，但是依赖于软件完成任务。软件在执行时是一系列机器指令进行取值、存储、加法等操作。操作系统是最基本的计算机软件，用来管理内存位置、软件指令、输入输出和其他程序的运行调度。

计算机软件可以用于完成特定任务，如字处理、记账，也可以适用于较宽的范围，比如电子表格软件可以用来记账、试算、作图，而**计算机语言**则是用来编制新的软件的工具。

计算机语言根据其运行方式可以分为解释型和编译型两种，解释型语言逐句解释程序并逐句执行，编译型语言把整个程序编译为二进制代码后再执行。按照计算机语言的抽象程度区分，包括二进制形式的机器语言，仅能用在特定的硬件中；汇编语言，用与 CPU 指令对应的命令编写，主要用于底层功能；面向细节的通用语言，如 Pascal, C, Fortran, Lisp, Cobol, C++, Java 等，优点是通用性和可复用性。

更高级的计算机语言有 R、Matlab 等，提供了包括向量、矩阵等高级数据类型，代码与统计的数学公式相似，直接支持求和、向量、矩阵等运算，易写易读，用户不用自己实现如解线性方程、求特征根这样的基础操作, 但是这样的语言一般是解释执行的，执行效率难以改进，不利于使用循环或迭代算法。本书使用 R 作为配套的编程语言。

R 软件是一个统计计算软件, 同时也是一种计算机编程语言，与 S 语言基本兼容。S 语言是 Rick Becker, John Chambers 等人在贝尔实验室开发的一个进行数据分析和交互作图的计算机语言，可以使用向量、矩阵、对象等数据进行编程，功能强大，程序简单。R 是 GPL 授权的自由软件，最初由新西兰 Auckland 大学的 Ross Ihaka 和 Robert Gentleman 于 1997 年发布，现在由 R 核心团队开发，全世界的用户贡献了数千个软件包，功能涵盖了经典和现代统计方法的绝大部分，是世界上许多顶尖的统计学家进行统计研究和发表算法的工具。见 R 的网站: `http://www.r-project.org/`。我们将讲授 R 的基本使用，在算法示例和习题中使用 R 作为编程语言，并在讲到具体统计计算方法时提及 R 中有关的函数。

本书目的主要是要求学生掌握统计计算方法、理解统计计算思想，但是这些算法的正确、高效实现也是很重要的。我们设置了足够的习题让学生通过实际编程了解算法实现中的各种问题。

在进行程序设计时，应注意以下几点：

- 要对程序抱有怀疑态度。逐步验证每一个模块。
- 大的任务要分解为小的模块。对每个模块进行详细测试。象 S 这样的更高级语言的优点是许多模块已经由系统本身提供了，如矩阵计算、函数优化等。但即使这样也应进行测试，因为系统提供的功能有可能不适用于你的特殊情况。
- 测试为一系列测试问题，涵盖应用的不同情况。最开始用最简单的测试。测试也应该包括超出范围的问题，这时程序应该能正确地判别错误输入。
- 好的程序不应该是用了很多高明的技巧以至于别人很难看懂，而是越直接越好，这样只要出错误就是明显的错误。

- 要有程序文档。
- 只要对程序的每一模块都进行了详细验证和检查，就可以确信程序的正确性。

1.1.3 内容提要

本书包括基本的统计计算方法，如计算分布函数、分位数函数的一般方法，矩阵计算方法、最优化方法、随机数生成算法。另外还用较大篇幅讲述了随机模拟方法，包括随机模拟的基本思想、改进精度的方法、重要应用。最后，还介绍了完全由计算方法发展出来的统计方法，如 Bootstrap、EM 算法、MCMC 方法等。

第一章后面部分先介绍 R 语言的基础，由于篇幅所限仅包括数据类型、程序结构、函数的基础介绍，进一步的用法还要参考其他教材和 R 用户手册。然后，讲解误差的来源和分类以及避免和减少误差的方法，这对我们了解算法的局限并在算法实现时避免产生有缺陷的算法有重要意义。最后，介绍描述统计量计算和简单的统计图形用法。

第二章为随机数产生与检验，包括均匀随机数的产生方法和检验方法，非均匀随机数的各种生成方法，包括函数变换、舍选抽样、重要抽样等。关于随机向量和随机过程的产生方法也进行了简单介绍。

第三章为随机模拟方法。首先，用随机模拟积分作为例子，讲解随机模拟方法的基本思想，包括减小随机模拟误差的技术。后面介绍了离散随机事件模拟中随机服务系统模拟问题。随机模拟方法对新统计方法的研究比较也具有广泛应用，本章用一个例子演示了随机模拟在统计方法研究中的用法。然后，本章讲述了 Bootstrap 方法，这是完全利用随机模拟方法解决统计推断问题的一个具体示例。MCMC 是现代统计计算的重要工具，尤其是在贝叶斯建模中起到关键作用，本章讲解了部分理论基础和实际做法。序贯重要抽样方法也是现代统计计算中一类重要方法，本章做了介绍。

第四章针对分布函数和分位数函数等近似计算问题，介绍函数逼近的多项式方法、连分数表示，插值方法，样条函数，数值积分和数值微分。

在回归分析等线性模型、多元模型、函数数据分析的计算中广泛用到矩阵计算方法。第五章介绍统计计算中常用的矩阵方法，比如矩阵三角分解、正交三角分解、特征值分解、奇异值分解，广义特征值，广义逆等。

很多统计计算问题都会归结为一个最优化问题，即求解函数的无约束或有约束的最小值（最大值）点的问题，比如最大似然估计、非线性回归等。第六章首先给出最优化问题的一些理论基础，然后讨论无约束最优化的方法，再给出约束最优化的一些算法，并讨论了统计计算中的一些特定的优化问题，如最大似然估计、非参数回归、EM 算法。

本书的每一章后面附有习题，有一些是对教材中理论的进一步延伸讨论，有一些是程序编制。读者可以选作这些习题，加深对教材内容的理解，并获得实际编程锻炼。

1.2 R 软件基础

R 是一个用于统计计算、绘图和数据分析的自由软件，最初由新西兰 Auckland 大学的 Ross Ihaka 和 Robert Gentleman 于 1997 年发布，现在由 R 核心团队开发维护，全世界的用户都可以贡献软件包。R 软件使用 R 语言，R 语言直接提供了向量、矩阵、数据框、一般对象等高级数据结构，其语法简单易学，功能强大，尤其适用于开发新的统计算法, 用 R 语言编写程序就像写公式一样简单，也可以调用 C、C++、Fortran 等编译语言的代码实现附加功能或提高计算效率。R 软件通过附加软件包 (package) 的方式提供了各种经典的统计方法以及最新的统计方法，到本书结稿时在 R 的网站 `http://www.r-project.org/`上已经有九千多个软件包。R 软件已经成为全世界统计学家研究新算法和进行统计计算的首选软件之一，在生物、金融、医药、工农业等各行各业的数据分析中也获得了广泛应用。

R 语言可以看成是 S 语言的一个变种，S 语言是 Rick Becker, John Chambers 等人在贝尔实验室开发的一个用于交互数据分析和作图的软件，最初的版本开发于 1976-1980 年，后来又有改进（见 Becker and Chambers(1984)[13], Becker et al(1988)[14], Chambers and Hastie (1992)[15]）。

这一节讲述 R 软件的基础用法, 这样读者可以用 R 语言实现自己的算法, 本书的算法也用类似 R 语言的语法描述。R 软件的进一步使用需要读者自己阅读 R 手册和 R 软件的相关书籍。

1.2.1 向量

R 软件可以在图形界面内以命令方式交互运行，也可以整体地运行存放在扩展名为`.r` 的程序文件中的代码。如果有某个 R 程序存放在当前目录中的 `mysrc.r` 文件中，如下命令可以运行这样的程序文件：

```
source('mysrc.r')
```

R 语言最基本的数据类型是**向量** (vector)，这是 R 的**对象**中最简单的一种。标量（单个的数值或字符串）可以看成是长度为 1 的向量。数值型的 R 向量基本可以看成是线性代数中的向量，但不区分行向量和列向量；与其它程序语言相比，R 向量相当于其它程序语言中的一维数组, 可以保存若干个相同基本类型的值，各个元素用向量名和序号（下标）访问。

R 中用函数 c() 来把若干个元素组合为一个向量，如

```
marks <- c(10, 6, 4, 7, 8)
```

定义了一个长度为 5、数据类型为数值 (numeric) 的向量，保存在名为 marks 的**变量**中。

R 中把向量等数据保存在变量中，如上例的 marks，用 <-为变量赋值。R 中的变量名可以由字母、数字、句点、下划线组成，变量名的第一个字符只能是字母或句点，如果以句点开头必须有第二个字符并且第二个字符不能是数字, 变量名长度不限。变量名是大小写区分的，所以 Amap 和 amap 是不同的名字。

为了访问向量 x 的第 3 个元素，只要用 x[3] 这种方法，下标是从 1 开始计数的。也可以直接修改一个元素，如

```
marks[3] <- 0
```

在 R 命令行环境中输入变量名然后按 Enter 键可以查看变量的值。如果在程序文件中要显示变量值 x 的值，应当使用 print(x)。更一般的输出可以用 cat 函数显示文本和数据值，如

```
cat(' 第三个分数 =', marks[3], '\n')
```

其中'\n' 表示换行。cat 的各项输出自动用空格分隔，在 cat 中加 sep='' 选项表示各项之间不分隔。

可以用 “开始值: 结束值” 的方法定义一个由连续整数值组成的向量，比如 11:13 为向量 $(11, 12, 13)$。用 rep(x, times) 可以把向量 x 重复 times 次，比如 rep(c(1,3), 2) 结果为向量 $(1, 3, 1, 3)$。用 rep(x, each=...) 可以把向量 x 的每个元素重复 each 次, 如 rep(c(1,3), each=2) 结果为向量 $(1, 1, 3, 3)$。

如果一个文本文件中用空格和换行分隔保存了多个数值，用函数 scan 可以把这些数值读入到一个数值型向量中，如

```
x <- scan('mydata.txt')
```

(其中 mydata.txt 是当前目录下的一个文本文件，保存了用空格和换行分隔的数值)。

R 的基本数据类型包括数值型 (numeric)、字符型 (character)、逻辑型 (logical)、复数型 (complex)。为了定义一个指定长度、元素初始化为 0 数值型向量，方法如

```
x <- numeric(4)
```

则变量 x 保存了由 4 个零组成的数值型向量。

字符型常量可以用两个单撇号或两个双撇号界定，如

```
tnames <- c(' 王思明', "John Kennedy")
```

R 的逻辑型只有真值 TRUE、假值 FALSE 和缺失值 NA 三个不同取值。

复数型常量写成如 1.2 - 3.3i 这样的形式，虚数单位 i 可以写成 1i。

和 C、C++、Fortran、Java 这些严格类型的程序设计语言不同，R 的变量不需要事先声明数据类型。为了动态地获知某个向量 x 当前所保存的基本数据类型，可以调用函数 mode(x)，结果为字符串'numeric'、'character'、'logical' 或 'complex'。基本类型之间可以用 as.xxx 类的函数进行转换，比如 as.numeric(c(TRUE, FALSE)) 结果为数值型向量 $(1, 0)$。

向量元素中用 NA 表示缺失值，如

```
x <- c(1, NA, 3)
s <- c(' 王思明', NA, 'John Kennedy')
```

但是没有上下文的一个 NA 被认为是逻辑型的。

向量是 R 对象中的一种。R 对象用**属性**用来区分不同数据类型或者对数据附加额外描述信息，基本类型 (mode) 就是所有 R 对象都有的基本属性，另一个基本属性是长度 (length)，用函数 length(x) 可以返回向量 x 的元素个数。

为了访问 R 向量的子集，除了一个元素可以用 "变量名 [下标]" 的格式访问之外，还可以用一个正整数向量作为下标，比如 marks[c(1,5)] 结果为 $(10, 8)$,

```
marks[3:5] <- 0
```

把 marks 的第 3 到第 5 号元素赋值为零。

向量下标还可以取为负整数或负整数向量，如 (11:15)[-5] 是 (11:15) 去掉了第 5 个元素之后的结果 $(11, 12, 13, 14)$, (11:15)[-c(1,5)] 是 (11:15) 去掉了第 1 和第 5 号元素之后的结果 $(12, 13, 14)$。

R 向量可以定义元素名并使用元素名作为下标，比如

```
marks <- c(' 李明'=10, ' 张红艺'=6, ' 王思明'=4,
           ' 张聪'=7, ' 刘颖'=8)
```

则可以用 marks[' 张聪'] 得到元素值 7，用 marks[c(' 李明', ' 张聪')] 得到结果 $(10, 7)$。

所有元素的元素名组成一个字符型向量，这是向量的对象属性之一。可以用 `names(x)` 取得向量 `x` 的元素名向量，也可以用如

```
names(marks) <- c(' 李明', ' 张红艺', ' 王思明', ' 张聪', ' 刘颖')
```

这样的方法定义或修改元素名属性。注意，这里赋值符号左边的 `names(marks)` 是对象属性的一种表示方法，不能看成是给一个函数结果赋值。

R 程序允许自动续行。只要前一行程序明显地没有完成，就可以直接拆分到下一行，没有配对的括号是前一行程序没有完成的一种常见情况。在一行的开头或中间以 # 号开始注释，# 以及该行的后续内容都作为注释。另外，两个语句可以用分号分隔后写在同一行中，比如

```
x <- 1; y < 2; z <- x+y
```

因子 (factor) 是一种特殊的向量：元素只能在有限个 “水平” 中取值，元素值编码为正整数值保存，可以用来表示分类变量的观测值。例如

```
fac <- factor(c(' 男', ' 男', ' 女', ' 男'))
```

生成一个水平为 “男” 和 “女” 的因子,用 `levels(fac)` 可以查看其各个水平,用 `as.numeric(fac)` 可以查看其编码值，此例为 $(1, 1, 2, 1)$; 用 `as.character(fac)` 可以转换为字符型。

1.2.2 向量运算

R 的算术运算符为

```
  +  -  *  /   %%   %/%   ^
```

分别表示加、减、乘、除、除法余数、整除、乘方。算术运算的优先级为先乘方、再乘除、最后加减，可以用圆括号来改变运算次序。有缺失值参加的四则运算结果还是缺失值。

向量可以和一个标量进行算术运算，结果为向量的每个元素与标量进行计算得到的新向量。如 `(11:14) + 100` 的结果为向量 $(111, 112, 113, 114)$，`2^(0:4)` 的结果为向量 $(1, 2, 4, 8, 16)$。

两个长度相同的向量进行算术运算，结果为对应元素进行算术运算后得到的新向量。比如 `(11:14) - (1:4)` 结果为向量 $(10, 10, 10, 10)$。

如果两个向量长度不同，但较大长度是较小长度的整数倍，则这两个向量也可以进行算术运算，在运算时把短的一个循环重复使用。比如，`(11:14) + c(100, 200)` 结果为向量 $(111, 212, 113, 214)$。

在 R 中，允许对一个向量调用数学中的一元函数，函数输出结果为每个元素取函数值后组成的新向量，比如 `sqrt(c(1,4,9))` 的结果为向量 $(1, 2, 3)$。类似的函数还有 `abs(x)` ($|x|$), `exp(x)` (e^x), `log(x)` ($\ln x$), `log10(x)` ($\lg x$), `sin(x)`, `cos(x)`, `tan(x)`, `asin(x)`(反正弦), `acos(x)`(反余弦), `atan(x)`(反正切), `atan2(y,x)`。`atan2(y,x)` 函数求平面直角坐标系中原点到点 (x, y) 的向量与 x 轴的夹角，对第 I、II 象限和 x 轴上的点，结果取值于 $[0, \pi]$，对第 III、IV 象限的点，结果取值于 $(-\pi, 0)$。

R 的比较运算符为

```
==  !=  <  <=  >  >=
```

可以表示两个标量比较，两个等长向量比较，或两个长度不等但是长度为倍数的两个向量比较，结果是对应元素比较的结果组成的逻辑型向量。如 `(1:4) > (4:1)` 结果为逻辑型向量 `(FALSE, FALSE, TRUE, TRUE)`, `(1:3) > 1.5` 的结果为逻辑型向量 `(FALSE, TRUE, TRUE)`, `(1:4) > (3:2)` 结果为逻辑型向量 `(FALSE, FALSE, FALSE, TRUE)`。

R 用`%in%` 运算符表示元素 “属于” 集合的判断，如果 `a` 是一个标量，`A` 是一个向量，把 `A` 看成一个集合（集合的元素就是 `A` 的各个分量），`a %in% A` 的结果就是 $a \in A$ 是否成立的结果。如果 `B` 是一个向量，`B %in% A` 就是对 `B` 的每个元素分别判断是否属于 `A` 的结果。比如, `c(2,-2) %in% (1:3)` 结果为 `(TRUE, FALSE)`。为了判断向量 `y` 的所有元素都属于向量 `x` 是否成立，可以用 `setdiff(y,x)` 求集合 y 减去集合 x 的差集，如果 `length(setdiff(y,x))==0` 为真则说明向量 `y` 的所有元素都属于向量 `x`。集合运算函数还有 `union(x,y)`(并集)、`intersect(x,y)`(交集)、`setequal(x,y)`(相同集合)、`is.element(el, set)`(属于判断)。

R 定义了如下的逻辑运算符：

```
&  |  !
```

其中& 表示逻辑与（同时为真才为真），| 表示逻辑或（任一为真则为真），! 表示逻辑非（真变成假，假变成真），两个逻辑向量之间可以进行逻辑运算，含义为对应元素的运算。一般用来把比较运算连接起来得到复杂条件，比如

```
(age >= 18) & (age <= 59)
```

表示年龄在 18 和 59 之间（含 18 和 59），

```
sex==' 女' | age <= 3
```

表示妇女或年龄在 3 岁以下（含 3 岁），

```
!((age >= 18) & (age <= 59))
```

表示年龄在 18 岁以下或 59 岁以上（不含 18 和 59）。

为了表示所有元素都满足一个条件，用函数 all(条件)，如 `all(age >= 18)` 表示向量 `age` 中所有年龄都在 18 岁以上。类似地，函数 any(条件) 表示向量中有任一元素满足条件。

R 中比较和逻辑运算的一个重要作用是按条件挑选向量元素子集。比如，要挑选向量 `marks` 中分数在 5 分以下（含 5 分）的元素，可以用如

```
marks[marks <= 5]
```

这是因为 R 的向量下标允许取为和向量长度相同的一个逻辑向量。这样取子集可能取到空集，结果表示为 `numeric(0)`，即长度为零的数值型向量。

1.2.3 矩阵

R 语言支持矩阵 (matrix) 和数组 (array) 类型, 矩阵是数组的特例。数组有一个属性 `dim`, 比如

```
arr <- array(1:24, dim=c(2,3,4))
```

定义了一个数组，元素分别为 1 到 24，每个元素用 3 个下标访问，第一个下标可取 1 和 2, 第二个下标可取 1, 2, 3, 第三个下标可取 1, 2, 3, 4。数组元素在填入各下标位置时，第一下标变化最快，第三下标变化最慢，这种排序叫做 FORTRAN 次序或列次序。比如，`arr` 中元素的次序为 `arr[1,1,1]`, `arr[2,1,1]`, `arr[1,2,1]`, `arr[2,2,1]`, ..., `arr[2,3,4]`。

矩阵是 dim 属性为两个元素的向量的数组，矩阵元素用两个下标访问，第一下标可以看作是行下标，第二下标可以看作列下标，元素按列优先次序存储。因为矩阵是最常用的数组，所以单独提供一个 `matrix` 函数用来定义矩阵，如

```
M <- matrix(1:6, nrow=2, ncol=3)
```

显示 `M` 的结果为

```
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

为了让矩阵元素按行填入，定义时可指定 `byrow=TRUE`，如

```
M <- matrix(1:6, nrow=2, ncol=3, byrow=TRUE)
```

显示 `M` 的结果为

```
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
```

单个数组元素用两个下标访问，如上述 M 的第 2 行第 3 列元素用 `M[2,3]` 访问。

可以取出 `M` 的一行或一列组成一个向量，格式如 `M[1,]`, 表示 `M` 的第一行组成的向量（不再是矩阵，所以不区分行向量和列向量），`M[,2]` 表示 `M` 的第二列组成的向量。

可以取出若干行或若干列组成子矩阵，比如 `M[,2:3]` 取出 `M` 的第 2 列和第 3 列组成的 2×2 子矩阵。取出子矩阵时如果仅有一行或仅有一列，子矩阵会退化成 R 向量而不再有维数属性，可以用 `M[,1,drop=FALSE]` 这样的方法要求取子集时数组维数不变。

用函数 `cbind` 可以把一个向量转换为列向量（列数等于 1 的矩阵），或者把若干个向量、矩阵横向合并为一个矩阵。用函数 `rbind` 可以转换行向量或进行纵向合并。

矩阵有两维（行维、列维），每一维都可以定义维名当作下标使用。对上述矩阵 `M`，如下命令定义了矩阵列名：

```
colnames(M) <- c('X1', 'X2', 'X3')
```

这样，M 的第二列可以用 M[,'X2'] 访问，M 的第一列和第二列组成的子矩阵可以用 M[,c('X1', 'X3')] 访问。类似地可以定义行名并用行名代替行下标：

```
rownames(M) <- c('John', 'Mary')
```

这时 M 显示如下：

```
     X1 X2 X3
John  1  3  5
Mary  2  4  6
```

列名和行名没有命名规则限制，但不应有重复值以免不能唯一地标识列和行。

矩阵首先是二维数组，形状相同的矩阵可以进行对应元素之间的四则运算，运算符与向量的四则运算符相同。比如，设

```
M2 <- rbind(c(1,-1,1), c(-1,1,1))
```

则

```
M + M2;  M - M2; M * M2;  M / M2;  M ^ M2
```

分别对 M 和 M2 的对应元素作加法、减法、乘法、除法、乘方，结果是形状相同的矩阵。对于加法和减法，结果与线性代数中两个矩阵相加和相减是一致的。

矩阵可以和标量作四则运算，结果是矩阵的每个元素与该标量进行四则运算。如果是矩阵与标量相乘，结果与线性代数中数乘结果相同。

两个矩阵相乘用%*% 运算符表示。比如

```
M %*% t(M1)
```

表示矩阵 M 左乘矩阵 M1 的转置 (函数 t 表示矩阵转置)。另外，crossprod(A) 表示 $A'A$, crossprod(A, B) 表示 $A'B$。

如果 A 是可逆方阵，用 solve(A) 求逆矩阵 A^{-1}。对线性方程组 $Ax = b$，用 solve(A, b) 可以求出线性方程组的解 x。

数据框 (data frame) 是类似于矩阵的数据类型，但是它允许同时保存数值型、字符型和因子型数据，每列的类型必须相同。数据框每列必须有变量名，用 names(df) 访问数据框

df 的各个变量名。统计数据经常以数据框的形式保存。用函数 read.csv 可以把 CSV(逗号分隔) 格式的文件转换为 R 数据框，用函数 write.csv 可以把 R 数据框保存为 CSV 格式的文件。

1.2.4 分支和循环

R 语言因为内置了向量、矩阵这样的高级数据类型，所以编程比较容易。很多其它语言中必须用分支、循环结构解决的问题在 R 语言中可以用更简洁的代码处理。

R 中用

if(条件) 语句 1 **else** 语句 2

处理分支结构，其中“条件”应该是一个逻辑型标量。else 部分可省略, 但是有 else 部分时两部分必须在同一个语句中。例如

```
if(x>0) y <- 1  else y <- 0
```

也可写成

```
if(x>0){
  y <- 1
} else {
  y <- 0
}
```

上例中使用了**复合语句**。R 中可以用左右大括号把若干个语句组合成一个复合语句，在程序中当作一个语句使用。

if 结构中的条件必须是标量。在上述问题中如果 x 是一个向量，可以改用如下的逻辑下标的方法：

```
y <- numeric(length(x))
y[x>0] <- 1
```

这样可以根据 x 元素的不同取值而对 y 的相应元素赋值。

R 语言中用 for 结构进行对向量下标或向量元素进行遍历（循环），格式为

for(循环变量 **in** 遍历向量**){**
  循环体语句
  ……
}

比如，下例中对向量 x 的下标循环以计算元素总和：

```
x <- 11:15
s <- 0
for(k in seq(along=x)){
  s <- s + x[k]
  cat('k=', k, 'x[k]=', x[k], 's=', s, '\n')
}
```

其中 seq(along=x) 获得向量 x 的下标向量（本例中为 1:5）。使用 seq(along=x) 而不是 1:length(x), 是因为 x 可以是零长度的，这时 1:length(x) 为 1:0, 即 1 和 0 两个元素，是错误的。还可以直接对 x 的元素循环，如

```
x <- 11:15
s <- 0
for(item in x){
  s <- s + item
  cat('item=', item, 's=', s, '\n')
}
```

在不能预知循环次数时，可以使用**当型循环**，格式为

while (循环继续条件 **)** {
  循环体语句
  ……
}

当循环继续条件成立时反复执行循环体语句，直到条件不成立时才不再执行。如果一开始循环继续条件就不成立，就一次也不执行循环体语句。

在计算机算法中还有所谓**直到型循环**，R 没有直接提供直到型循环的语法结构，在本书后续的算法描述中有时用如下的直到型结构:

until (循环退出条件) {
    循环体语句
    ……
}

执行循环直到循环退出条件成立时才退出循环，循环至少执行一次。R 中可以用如下方法模仿直到型循环:

repeat {
    循环体语句
    ……
    if (循环退出条件) **break**
}

其中 `repeat` 是无条件循环语句，`break` 跳出一重循环。

在本书的算法描述中，使用了类似于 R 语言语法的**伪代码**, 伪代码与程序类似，但允许使用描述性的操作说明。伪代码中主要用到 `if ... else ...`, `for` 循环, `while` 循环，`until` 循环等控制结构。

在 R 中，`for`、`while`、`repeat` 循环效率较低，比用 FORTRAN、C 等编译型语言的计算速度可能慢几个数量级。R 中的循环常常是可以设法避免的。比如，`sum(x)` 可以直接得到 `x` 的元素和。类似的函数还有 `mean(x)` 求平均值，`prod(x)` 求所有元素的乘积，`min(x)` 求最小值，`max(x)` 求最大值，`sd(x)` 求样本标准差，`cumsum(x)` 计算元素累加和（包括各中间结果），`cumprod(x)` 计算元素累乘积，等等。

R 的 `apply`、`lapply`、`sapply`、 `mapply`、 `vapply`、 `replicate`、 `tapply`、`rapply` 等函数隐含了循环。函数 `apply` 可以对矩阵的各行或各列循环处理，比如，`apply(M, 1, sum)` 表示对 `M` 的各行求和，而 `apply(M, 2, mean)` 表示对 `M` 的各列求平均。函数 `lapply(X,FUN, ...)` 可以把函数 `FUN` 分别作用于 `X` 的每一个元素 (比如数据框的每个变量)，输出一个包含这些作用结果的列表，`...` 是 `FUN` 所需要的其它自变量。`sapply(X,FUN,...)` 与 `lapply(X,FUN...)` 类似但尽可能把结果转换为向量或数组。`vapply` 与 `sapply` 作用类似，但需要自己指定输出类型。`mapply(FUN, ...)` 是把 `sapply` 推广到 `FUN` 为多元函数的情况。`replicate(n, expr)` 经常用于重复 `n` 次模拟，`expr` 是调用一次模拟的表达式。`rapply` 可以把函数作用到嵌套列表的最底层。

R 中提供了若干个函数可以进行快速计算，如 `fft` 计算离散傅立叶变换，`convolve` 计算离散卷积，`filter` 计算滤波或自回归迭代。R 这样的函数很多，需要用户能够根据函数的帮助信息，自己构造一些简单可手工验证算例，快速掌握这些函数的使用。

对于必须使用循环的情况，如果需要重复的次数不多，用 R 的循环不会造成效率损失；如果重复次数很多，可以考虑把重复很多的循环改为 C 代码或 C++ 代码，R 的 Rcpp 软件包支持在 R 代码内嵌入 C 代码和 C++ 代码。

1.2.5 函数

R 内置了许多函数，比如，`sin(x)` 可以计算向量 `x` 的每个元素的正弦值，`seq(a,b)` 可以生成从 `a` 到 `b` 的等差数列，`sum(x)` 可以计算 `x` 的所有元素的总和。

R 的函数在调用时可以按照自变量位置调用，如 `seq(2,5)` 结果为 (2,3,4,5), 也可以在调用时指定自变量名，这时自变量次序不再重要，如 `seq(to=5, from=2)` 结果还是 (2,3,4,5)；还可以前一部分按自变量位置对应，后一部分指定自变量名，如 `seq(2, length=4)` 结果仍为 (2,3,4,5)。

模块化程序设计是保证软件正确、可复用、易升级的重要方法，而自定义函数是模块化设计的主要组成部分。在 R 中用如下格式定义一个新的函数：

函数名 <- **function(**形式参数表**){**
    语句
    .....
    返回值表达式
}

其中函数体内最后一个表达式的值是函数的返回值。形式参数表可以为空，表示不需要自变量，但是括号不能省略。形式参数表可以是用逗号分开的形式参数名，并可以带有用等于号给出的缺省值。例如

```
demean <- function(x, xbar=mean(x)){
  x - xbar
}
```

调用如 `demean(c(1,2,3))`, 则缺省参数 `xbar` 用给出的表达式计算，返回值为 (-1,0,1)。还可以调用如 `demean(c(1,2,3), xbar=1)` 或 `demean(c(1,2,3), 1)`，结果为 (0,1,2)。

函数只能返回一个表达式的结果，可以是标量、向量、矩阵、数据框或其它更复杂的类型。为了能够同时返回多个结果，可以把这些结果组合成为一个**列表** (list) 类型。列表定义如

```
list(coefficients=b, Fvalue=F, pvalue=pv)
```

则此列表包含了回归系数向量、检验的 F 统计量和相应的 p 值。设上述列表保存在列表变量 `li` 中，为了访问 `li` 的 `pvalue` 元素，可以用 `li[['pvalue']]`、`li$pvalue` 或 `li[[3]]` 的格式。用 `names(li)` 查询或修改列表的元素名。为了删除列表中某项，只要将其赋值为特殊的 `NULL` 值，表示不存在的值，`NULL` 与 `NA` 不同，`NA` 表示存在但是缺失的值。

R 自定义函数内，形式参数是局部的，即形式参数和函数定义外其它变量重名时不会有冲突。在函数定义内赋值变量是局部的，即使函数定义外部有同名变量也不会给外部的同名变量赋值。然而，如果在函数定义内读取某个变量的值，此变量的值在函数内无定义但是在它的外部环境中有定义，运行时可以读取外部定义的值，这是一个需要小心的特点，因为在较长的函数定义中可能忘记给变量赋值而使用了外部环境中同名变量的值。

1.3 误差

1.3.1 误差的种类

统计计算的算法要得到正确的结果，就需要尽可能减少误差。统计问题中的误差有模型误差、实验误差和数值计算误差，在统计计算研究中主要解决的是如何减少数值计算误差的问题。

统计计算的算法通常是用来求解某种统计模型。任何用来解决实际问题的数学模型都或多或少地简化了实际问题，忽略掉一些细节，从而**模型误差**不可避免。如果模型不合适，其它误差控制得再完美，问题也不能得到解决；更糟的是，良好的计算结果会给使用者以错误的信心。比如，我们使用的回归模型要求观测是独立的，而实际数据观测有不可忽略的序列相关性，尽管我们用软件算出了很完美的结果，这个结果也是错误的。我们应当仔细选择模型，尽可能减少模型误差。

建立统计模型所需的数据来自实验、观测、抽样调查等过程，在这样的过程中会出现**实验误差**，包括随机误差、系统误差、过失误差。

随机误差是试验过程中由一系列随机因素引起的不易控制的误差，可以通过多次重复试验或改进模型设计来减小随机误差。随机误差可能来自物理量本身的波动，比如测量风速，就是在测量一个随时变化的物理量，不可避免地受到随机误差影响。随机误差可能来自不可控制的随机因素影响，比如，在用雷达测量飞机的方位和速度时，可能受到地磁、气温、地形的影响。由于测量仪器精度的限制也会产生随机误差，比如用最小刻度是 1 度的温度计测

量温度，只能把不足 1 度的值四舍五入或者估计小数点后一位数字。随机误差也可能来自特定条件下才发生的程序错误。

系统误差是多次测量持续偏高或偏低的误差，多次重复测量不能消除或减少系统误差。系统误差可能来自仪器本身的误差，比如用不锈钢直尺测量家具高度，直尺本身在温度不同时长度有细微变化。系统误差也可能来自仪器使用不当，比如用天平测量质量时天平没有配准。当发现有系统误差时，必须找出引起误差的原因并消除。

在记录实验数据时由于人的过失可以导致误差发生, 这样的误差称为**过失误差**。比如，在记录仪表（如水表、电表）的读数时看错数字，在记录数值时写错小数点位置，在上传数据时报告了过时的或错误的数据，等等。统计数据分析必须甄别并改正这样的过失误差，否则会对分析结果产生严重影响。在使用计算机软件处理数据时程序设计的质量问题也会导致误差发生。比如，当输入条件不满足模型要求而程序中未进行检查时，可能给出错误的结果。

1.3.2 数值计算误差

数值计算误差是用电子计算机进行数据存储和计算时产生的误差，设计算法时必须了解并尽可能避免这种误差。

设 A 为结果的精确值，a 为计算得到的近似值，则 $\Delta = a - A$ 称为**绝对误差**, 简称误差。$\delta = \frac{a-A}{A} = \frac{\Delta}{A}$ 称为**相对误差**。相对误差没有量纲，常常用百分数表示。绝对误差和相对误差常常仅考虑其绝对值。实际中如果能估计绝对误差和相对误差的取值区间或绝对值的上限，则可以确定误差大小。如果绝对误差 Δ 的绝对值上限估计为 $\tilde{\Delta}$, 则相对误差的绝对值上限可以用 $\delta = \frac{\tilde{\Delta}}{a}$ 估计，因为 A 的真实值一般是未知的。有 p 位有效数字的一个十进制近似数的相对误差约为 5×10^{-p} （见习题5）。

没有数值计算经验的读者往往会有一个错误认识，认为计算机得到的结果总是准确的、可信的。即使不考虑模型误差、随机误差和系统误差的影响，用计算机进行数值计算也不可避免地受到误差影响，只要我们把误差控制在可接受的范围就可以了。

为了解数值计算误差，我们首先需要了解计算机中数值的存储方式。计算机中的数都用二进制表示，包括定点表示和浮点表示两种。定点数主要用于保存整数，是精确表示的，定点数的四则运算可以得到精确结果，但整数除法需要特别注意余数问题，另外定点表示的整数范围有限，比如用 32 个二进制位可以表示 $-2^{31} \sim 2^{31} - 1$ 之间的整数（约 10 位有效数字），定点数的计算结果有可能溢出，即超出能表示的整数范围。

数值计算时主要使用浮点表示, 比如用 64 个二进制位能表示绝对值在 $10^{-308} \sim 10^{308}$ 之间的实数，有效数字大约有 $16 \sim 17$ 位。二进制浮点数的表示包括 (S, E, F) 三个部分，即正负号、指数部分、小数部分。小数部分的位数 d 决定了精度。一般存储 $2^{-1} \le F < 1$, 这样 F

中首位一般是 1，有些计算机不保存这一位。例如，二进制数 $101.101_2 = 0.101101_2 \times 2^3$ 可以表示为 $(+, +11_2, .101101_2)$。

由于 E 和 F 两部分的位数限制，二进制浮点数只有有限个，这些数的集合记为 $\mathcal{F}$，在这个集合中可以制定一个次序，(S, E, F) 的下一个数是 $(S, E, F + 2^{-d})$。对于实数域 $\mathbb{R}$ 中的实数 x，当 $|x|$ 超出浮点数范围时，我们不能表示；当 $|x|$ 没有超出浮点数范围时，一般也只能用 $\mathcal{F}$ 中的数来近似表示。比如，十进制数 0.1 如果表示成二进制数，是一个无限循环小数 $0.000\dot{1}100\dot{0}_2$, 只能存储其中有限位。在近似表示时对于不能保存的位数计算机中可能用了舍入法或者截断法处理, 这样造成的误差叫做**舍入误差**。

设实数 z 的浮点表示为 $\mathrm{fl}(z)$，则绝对舍入误差满足

$$|\mathrm{fl}(z) - z| \leq U|z|, \forall z.$$

其中 U 称为机器单位 (machine unit)，fl 用截断近似时 $U = 2^{-(d-1)}$, fl 用四舍五入时 $U = 2^{-d}$。一个浮点数 z 用浮点表示 $\mathrm{fl}(z)$ 后误差范围如下

$$\mathrm{fl}(z) = z(1 + u),\ |u| \leq U.$$

和 U 类似的一个数是机器 ε_{m}，ε_{m} 是使得 $1 + \varepsilon_{\mathrm{m}}$ 和 1 的表示不相同的最小正浮点数。U 和 ε_{m} 很相近，可以互相替代使用。典型的双精度计算 ε_{m} 约为 10^{-16} 数量级，而单精度的相应值与双精度相差约 10^9 倍, 即单精度计算与双精度计算的精度相差约 9 位有效数字。使用双精度实数可以减少表示误差和计算误差，但是不能消除这两种误差，所以不能盲目相信使用了双精度后就能得到正确的数值计算。

在 R 软件中用变量`.Machine$double.eps` 保存双精度计算的机器 ε_{m} 的值。

浮点数的四则运算不符合结合律，计算结果与计算次序有关。不同的算法可能导致不同的计算误差，应该尽可能选用计算精度高的数学公式，另外在设计算法时需要注意避免一些损失精度的做法。

例如，计算

$$\sum_{n=1}^{1000} \frac{1}{n(n+1)}$$

可以直接累加计算，造成很大的累积误差。只要把公式变换成

$$\sum_{n=1}^{1000} \frac{1}{n(n+1)} = \sum_{n=1}^{1000} \left(\frac{1}{n} - \frac{1}{n+1} \right) = 1 - \frac{1}{1001}$$

就只要计算一个除法和一个减法。

例 1.3.1. 计算多项式

$$P_n(x) = a_0 + a_1x_1 + \cdots + a_nx^n$$

时，直接计算，a_kx^k 需要计算 k 次乘法，总共需要计算 $1+2+\cdots+n = \frac{1}{2}n(n+1)$ 次乘法和 n 次加法, 我们称这样的算法需要 $O(n^2)$ 次浮点运算。如果用如下递推算法（秦九韶法）

$u_0 \leftarrow a_n$
for(k **in** $1:n$) {
  $u_k \leftarrow u_{k-1} \cdot x + a_{n-k}$
}
输出 u_n

就只需要 n 次乘法和 n 次加法, 只需要 $O(n)$ 次浮点运算，在提高了效率的同时也会提高计算精度。 □

在进行浮点数加减时，两个绝对值相差很大的数的加减严重损失精度。设 $|a| \gg |b|$($\gg$ 表示“远大于”), 则 $a+b$ 会被舍入为 a。比如，计算

$$0.1234 + 0.00001234$$

如果仅允许保留 4 位有效数字，则结果为 0.1234。为避免这样的问题，应该只对大小相近的数相加减，比如，可以把小的数组合后分别相加，再加起来。

两个相近数相减损失有效数字。如:

$$0.8522 - 0.8511 = 0.0011$$

有效数字个数从 4 位减低到 2 位。统计中如下的方差计算公式：

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2$$

就存在这样的问题。在对分布函数 $F(x)$ 计算右尾部概率 $1-F(x)$ 时，如果 x 很大则 $F(x)$ 很接近 1，计算 $1-F(x)$ 会造成结果有效数字损失。为此，统计软件中计算分布函数时常常可以直接计算右尾概率，以及概率的对数。

例 1.3.2. 逻辑斯蒂分布函数为 $F(x) = \frac{1}{1+e^{-x}}$, 当 x 很大时如果直接计算 $1 - \frac{1}{1+e^{-x}}$，就会出现有效位数损失；只要改用

$$1 - F(x) = \frac{1}{1+e^x}$$

计算，就可以避免两个相近数相减，提高精度。用 $1-\frac{1}{1+e^{-x}}$ 计算 $1-F(8)$ 结果为 $4.539786870239038\times10^{-05}$, 用 $\frac{1}{1+e^{x}}$ 计算 $1-F(8)$ 结果为 $4.539786870243439\times10^{-05}$，第一种计算公式损失了 4 位以上的有效数字。 □

例 1.3.3. 计算 $\sqrt{x^2+1}-|x|$，直接计算在 $|x|$ 很大时会有较大误差，改成 $1/(\sqrt{x^2+1}+|x|)$ 则不损失精度。 □

如果某函数参数相对变动在机器精度 U 附近时函数值也有变化，则函数值的计算结果被舍入误差严重影响，不能得到有效计算结果。

多次迭代相加或相乘造成舍入误差积累。比如,把 0.1 累加一千万次,结果为 1000000.00000008917, 在第 15 个有效位上出现了误差。

浮点数做乘法运算时，结果有效位主要取决于精度较低的数。做除法运算时，如果分母绝对值很小则会产生较大的绝对误差。

在比较两个浮点数是否相等时，因为浮点数表示和计算的不精确性，程序中不应该直接判断两个浮点数是否相等，而应该判断两者差的绝对值足够是否小。

因为计算机浮点数的不精确性，所以一定要注意浮点算法的局限。要了解能保存的最大和最小实数。例如，单精度浮点数的范围有时不够用。超过最大值会发生**向上溢出**, 绝对值小于最小正浮点数会发生**向下溢出**。向上溢出一般作为错误；向下溢出经常作为结果零。还要了解相对误差的范围，用机器精度 U 或 ε_m 表示。

例 1.3.4. 设一个班有 n 个人，则至少有两人的生日的月、日重合的概率为

$$p_n = 1-\frac{P_{365}^n}{365^n}$$

其中 $P_{365}^n = 365\times364\times\cdots\times(365-n+1)$。如果直接计算分式的分子和分母，则当 n 达到 121 时分母的计算溢出。只要改用如下计算次序就可以解决这个问题：

$$P_n = 1-\prod_{j=0}^{n-1}\frac{365-j}{365}$$

□

除了数值计算中浮点数表示和计算带来的误差，另一类更大的误差是**截断误差**。比如，计算机中的超越函数如 e^x 通常用级数逼近，而级数只能计算到有限项，舍去部分造成误差。在确定需要计算的项数时，机器精度 U 或 ε_{m} 给出了一个最大界限，任何使得结果相对变化小于 U 或 ε_{m} 的尾项都不需要再加进来。数值计算中方程求根、求最小值点经常使用迭代法求解，而迭代法也不能无限执行，一般预定一个精度，达到精度时计算停止，会造成误差。所以，即使是编程语言内置的函数或公认的程序包的结果也是有数值计算误差的。

1.3.3 随机误差的度量

随机误差来自于观测样本或随机模拟中的随机因素，一般随着样本量增大而减小，但随机误差是随机变量，不能给出严格误差界限，只能用概率方法描述随机误差大小。

为了方便讨论当样本量 n 趋于无穷时随机误差大小的变化规律，引入如下的 O_p 和 o_p 的记号。

$O_p(\cdot)$ 是概率论中表示依概率有界的记号。如果 $\{\xi_n\}$ 和 $\{\eta_n\}$ 是两个随机变量序列，满足

$$\lim_{M\to\infty}\sup_{n\geq 1} P\left(\left|\frac{\xi_n}{\eta_n}\right| > M\right) = 0$$

则称 $\{\frac{\xi_n}{\eta_n}\}$ **依概率有界**，记为 $\frac{\xi_n}{\eta_n} = O_p(1)$ 或 $\xi_n = O_p(\eta_n)$。

如果对 $\forall\delta > 0$ 都有

$$\lim_{n\to\infty} P\left(\left|\frac{\xi_n}{\eta_n}\right| > \delta\right) = 0$$

称 ξ_n/η_n **依概率趋于零**，记为 $\xi_n/\eta_n = o_p(1)$ 或 $\xi_n = o_p(\eta_n)$。

O_p 和 o_p 是我们用来估计随机误差幅度的有效工具，有如下性质:

(1) $o_p(1) = O_p(1)$;

(2) 如果 $\{\eta_n\}$ 是趋于零的常数列，则 $\xi_n = O_p(\eta_n)$ 时必有 $\xi_n = o_p(1)$。

(3) $o_p(1) + o_p(1) = o_p(1)$;

(4) $o_p(1) + O_p(1) = O_p(1)$;

(5) $o_p(1) \cdot o_p(1) = o_p(1)$;

(6) $o_p(1) \cdot O_p(1) = o_p(1)$;

(7) 如果 $\{\xi_n\}$ 以概率 1 趋于零，或 $\{\xi_n\}$ 是趋于零的非随机实数列，$\xi_n = o_p(1)$;

(8) 单个随机变量 $\xi = O_p(1)$;

(9) 如果随机变量序列 ξ_n 依分布收敛到分布 F 则 $\xi_n = O_p(1)$，且 $\xi_n + o_p(1)$ 仍依分布收敛到分布 F。

如果 ξ_n 满足中心极限定理

$$\frac{\xi_n - E\xi_n}{\sqrt{\mathrm{Var}(\xi_n)}} \xrightarrow{d} \mathrm{N}(0,1)$$

设 $n\mathrm{Var}(\xi_n)$ 有上界 D，则

$$\xi_n - E\xi_n = \frac{\sqrt{n\mathrm{Var}(\xi_n)}}{\sqrt{n}} \cdot O_p(1) = O_p\left(\frac{\sqrt{D}}{\sqrt{n}}\right) = o_p(1).$$

1.3.4 问题的适定性与算法稳定性

把算法粗略看成

$$\text{输出} = f(\text{输入})$$

问题的适定性可以看成是输入微小变化时输出变化的大小，如果输出连续地依赖于输入并且输入的微小变化引发的输出变化也是微小的，则问题是适定的。为了问题的适定性，定义**条件数**如下：

$$\frac{|f(\text{输入}+\delta) - \text{输出}|}{|\text{输出}|} = \text{条件数} \cdot \frac{|\delta|}{|\text{输入}|}.$$

即条件数是输出的相对变化与输入的相对变化的比值。当 $f(x)$ 为一元可微函数时，计算 $f(x)$ 的条件数可以用微分表示为

$$\kappa(f, x) = |xf'(x)/f(x)|.$$

条件数较小，比如 $\kappa < 10$ 的时候，可以设计算法给出问题的精确解。条件数很大的问题称为**病态问题** (ill-conditioned), 可能会有很大误差。条件数等于无穷或不存在的问题称为**不适定问题** (ill-posed)。

即使问题是适定的，在数值计算中由于浮点计算有限精度的影响结果也可能不稳定。稳定的算法应该是适定的，而且不受浮点数计算误差的影响。

考虑二次方程

$$z^2 - x_1 z + x_2 = 0, \ (x_1 > 0, x_2 > 0) \tag{1.1}$$

设其中 x_2 很小，这时两个根都是正根，求其中较小根 z_2。z_2 定义为

$$z_2 = z_2(x_1, x_2) = \frac{x_1 - \sqrt{x_1^2 - 4x_2}}{2} \tag{1.2}$$

在 x_2 变化时条件数

$$\begin{aligned} C =& \left| x_2 \frac{\partial z_2(x_1, x_2)}{\partial x_2} / z_2(x_1, x_2) \right| \\ =& x_2 \frac{1}{x_1^2 - 4x_2} \frac{2}{x_1 - \sqrt{x_1^2 - 4x_2}} = \frac{z_1}{z_1 - z_2} \end{aligned}$$

当 z_1 较大，z_2 很小时 C 接近 1，问题适定性很好。当判别式 $x_1^2-4x_2$ 很接近于零时两个根 z_1 和 z_2 差很小，条件数很大，这时 x_2 的一个微小变化可能引起根 z_2 的很大变化。

即使问题适定，不适当的算法也可能会造成结果不稳定。比如，即使(1.1)的判别式 $x_1^2-4x_2$ 不接近于零，两个根 z_1 和 z_2 差距很大，但是当 x_2 很小时公式(1.2)的分子中 x_1 和 $\sqrt{x_1^2-4x_2}$ 很接近，相减会造成精度损失，改用公式

$$z_2=\frac{2}{x_1+\sqrt{x_1^2-4x_2}} \tag{1.3}$$

则避免了相近数相减，提高了精度。例如，取 $x_1=1$, $x_2=0.5\varepsilon_m$, 则条件数 $C\approx 1$, $x_2+\delta=0.6\varepsilon_m$, 则公式(1.2)的计算结果相对变化为 0，而公式(1.3)的计算结果相对变化为 0.2，与 $C\delta$ 相等。

稳定性研究经常用倒推法：设精确结果应该为 $y=f(x)$, x 是输入，y 是输出。有误差的结果记作 $y^*=f^*(x)$, 设 y^* 等于某个输入 x^* 的精确结果 $f(x^*)$，则 $|x-x^*|$ 的大小代表了算法的稳定程度, 如果 $|x-x^*|$ 很小则称算法是倒推稳定的，$|x-x^*|$ 称为倒推误差。

1.4 描述统计量

1.4.1 总体和样本

统计的主要作用在于分析数据，发现数据中的规律。统计学是以概率论为数学基础的。我们先回顾一些基本概念。

假设一个变量的多个观测值 $y_1,y_2,\dots,y_n$ 可以看成某随机变量 Y 的 n 个独立观测，称 Y 为**总体**，称 $y_1,y_2,\dots,y_n$ 为总体 Y 的**简单随机样本** (简称**样本**)，称从样本计算得到的量为**统计量**。

对样本 $y_1,y_2,\dots,y_n$ 从小到大排序得到 $y_{(1)}\le y_{(2)}\le\cdots y_{(n)}$, 称为样本的**次序统计量**。

为了估计 $F(y)=P(Y\le y)$, 使用如下统计量

$$F_n(y)=\frac{\#(\{i:y_i\le y\})}{n}=\frac{1}{n}\sum_{i=1}^{n}I_{(-\infty,y]}(y_i)$$

其中 $\#(A)$ 表示集合 A 中元素的个数，$I_A(y)$ 是集合 A 的示性函数，当 $y\in A$ 时等于 1，当 $y\notin A$ 时等于 0。$F_n(y)$ 是样本统计量，也是 y 的函数，称 $F_n(y)$ 为**经验分布函数**。如果

$y_{(1)} < y_{(2)} < \cdots < y_{(n)}$，易见

$$F_n(y) = \begin{cases} 0 & \text{当 } y < y_{(1)} \\ \frac{1}{n} & \text{当 } y_{(1)} \le y < y_{(2)} \\ \cdots & \cdots\cdots \\ \frac{i-1}{n} & \text{当 } y_{(i-1)} \le y < y_{(i)} \\ \cdots & \cdots\cdots \\ \frac{n-1}{n} & \text{当 } y_{(n-1)} \le y < y_{(n)} \\ 1 & \text{当 } y \ge y_{(n)} \end{cases}$$

可见 $F_n(y)$ 是 y 的单调递增右连续阶梯函数，跳跃点为各 $y_{(i)}$ 处，跳跃高度为 $\frac{1}{n}$。如果有 $y_{(i)}$ 有 k 个相等的观测值，则在 $y_{(i)}$ 处跳跃高度为 $\frac{k}{n}$。由强大数律易见 $\lim_{n\to\infty} F_n(x) = F(x)$, a.s. $\forall x \in (-\infty, \infty)$, 即 $F_n(y)$ 是 $F(y)$ 的强相合估计。进一步地，$F_n(y)$ 是 $F(y)$ 的一致强相合估计。

事实上，$F_n(y)$ 也是一个真正的分布函数。设随机变量 $W \sim F_n(y)$, 则 W 服从离散分布，在 $\{y_1, y_2, \ldots, y_n\}$ 内取值，如果各 y_i 互不相同则 W 服从 $\{y_1, y_2, \ldots, y_n\}$ 上的离散均匀分布，$P(W = y_i) = \frac{1}{n}$, $i = 1, 2, \ldots, n$。如果 $\{y_1, y_2, \ldots, y_n\}$ 中有相同的观测值则其相应的取值概率是 $\frac{1}{n}$ 乘以重复次数。

下面给出**分位数函数**的定义。设随机变量 X 的分布函数 $F(x)$ 严格单调上升且连续，则存在反函数 $F^{-1}(p)$ $(0 < p < 1)$ 使得

$$F(F^{-1}(p)) = p, \ \forall 0 < p < 1.$$

称 $F^{-1}(p)$ 为 X 或 $F(x)$ 的分位数函数，称 $x_p = F^{-1}(p)$ 为 $F(x)$ 的 p 分位数。如果 $F(x)$ 没有反函数 (当 $F(x)$ 在某个区间等于常数时)，则只要 x_p 满足

$$P(X \le x_p) \ge p, \quad P(X \ge x_p) \ge 1 - p$$

就称 x_p 是 X(或 $F(x)$) 的 p 分位数。这样的 x_p 可能不唯一，取其中最小的一个，可得

$$x_p = \inf\{x : F(x) \ge p\}.$$

1.4.2 样本的描述统计量

在对数据建模分析之前，我们需要预先排除数据中的错误，了解数据的特点，考察数据是否符合模型假定的前提条件。这些前期准备工作统称为**探索性数据分析** (Exploratory Data Analysis, EDA)。

最常见的数据形式是一个变量的多次观测，或一组变量的多次观测。对单个变量，我们关心其分布情况；对多个变量，我们还关心变量之间的关系。

下面介绍用描述统计量来概括数据的方法。

随机变量主要有两种，离散型和连续型。离散型随机变量可以代表某种分类值，比如性别、行业、省份，也可以是离散取值的数值变量，比如 n 次独立重复试验的成功次数，一块电路板上的缺陷个数，等等。

对于取分类值的随机变量，EDA 的作用是考察其取值集合，计算取每个不同值的次数和比例，并纠正不一致的输入。比如，变量

```
sex <- factor(c(' 男', ' 女', ' 女', ' 男', ' 男生'))
```

保存了 5 个人的性别，用

```
table(sex)
```

可以求得每个不同值的取值次数，并且可以发现输入存在不一致性。

对于数值型的变量（包括离散取值和连续取值），我们主要关心其分布情况。如果是离散取值的，我们需要确定其可取值集合；如果是连续取值，我们需要了解其可取值的区间或集合。然后，我们可以从样本中计算描述分布特征的统计量，来描述变量分布。称这样的统计量为**描述统计量**。

作为基础，先给出估计分位数的方法。设 $y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)}$ 为样本的次序统计量，用

$$\hat{x}_p = y_{([np])}$$

可以估计 x_p。其中 $[x]$ 表示对实数 x 向下取整, 即小于等于 x 的最大整数。当 x_p 唯一时，$\hat{x}_p$ 是 x_p 的强相合估计。但是，当 n 不太大时，这样的估计方法有些粗略。改进 x_p 估计的想法是，因为 x_p 是 $F(y) = p$ 的解，可以用适当方法把阶梯函数 $F_n(y)$ 变成严格单调上升连续函数，设改进后 $\tilde{F}_n(y) \approx F(y)$，再解 $\tilde{F}_n(y) = p$ 得到 x_p 的估计。例如，令

$$\tilde{F}_n(x_{(i)}) = \frac{i - \frac{1}{3}}{n + \frac{1}{3}},\ i = 1, 2, \ldots, n$$

(如果有若干个 $x_{(i)}$ 取值相等，i 取最小下标) 并把相邻点用线段相连来定义 $\tilde{F}_n(x_{(i)})$，求 $\tilde{F}_n(x_{(i)}) = p$ 得到 x_p 的估计。详见 Hyndman and Fan(1996)[22]。R 语言中函数 `quantile(x)` 可以计算样本分位数，并提供了 Hyndman and Fan(1996)[22] 描述的 9 种不同的计算方法。

为了概括地描述数值型的随机变量分布，可以使用以下几类常用的数字特征：

一、位置特征量。假设 $F(x)$ 是一个分布函数，则 $\{F(x-\theta), \theta \in \mathbb{R}\}$ 构成了一族分布，其中 θ 是代表分布位置的数字特征。例如，$\mathrm{N}(\mu, \sigma^2)$ 分布中 μ 是代表分布位置的数字特征。

可以用来描述总体 Y 的分布位置的数字特征包括期望值 EY，中位数 m, 众数 d。

期望值 EY 是样本取值用取值概率或概率密度作为加权的加权平均。比如，在总共 10000 张售价 1 元的彩票中仅有 1 张是有 5000 元奖金的，则随机抽取一张的获利 Y 的期望为如下加权平均：

$$\begin{aligned} EY =& (5000-1) \times P(\text{获奖}) + (-1) \times P(\text{不获奖}) \\ =& (5000-1) \times \frac{1}{10000} + (-1) \times \frac{9999}{10000} = -0.5(\text{元}) \end{aligned}$$

对样本 $y_1, y_2, \ldots, y_n$，用样本平均数

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

估计 EY。在 R 语言中用 `mean(x)` 求平均值。

样本平均值受到极端值的影响很大。比如，某企业有 1 名经理和 100 名员工，经理月薪 100000 元，其他员工月薪 1000 元，则 101 位雇员的平均工资为

$$\bar{y} = \frac{1}{101}(100000 + 100 \times 1000) \approx 1980.2$$

并不能真实反映工资的一般情况。

如果我们需要递推地计算样本平均值，记 $\bar{y}_k = \frac{1}{k} \sum_{i=1}^{k} y_i$，可以用如下算法：

$\bar{y}_1 = y_1$
for (k **in** 2:n) {
    $\bar{y}_k = \frac{k-1}{k} \bar{y}_{k-1} + \frac{1}{k} y_k$
}

中位数 m 是满足

$$P(Y \le m) \ge \frac{1}{2}, \quad P(Y \ge m) \ge \frac{1}{2}$$

的数, 即 Y 的 $\frac{1}{2}$ 分位数。当 n 为奇数时令 $\hat{m} = y_{(\frac{n+1}{2})}$, 即从小到大排序后中间一个的值，当 n 为偶数时令 $\hat{m} = \frac{1}{2}(y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)})$, 即从小到大排序后中间两个的平均值。称这样得到的 $\hat{m}$ 为**样本中位数**，用来估计总体中位数 m。在 R 语言中用 `median(x)` 求样本中位数。

总体众数 d 是 Y 的分布密度的最大值点，或 Y 的概率分布中概率值最大的取值点，可以用样本值中出现最多的一个数来估计。如果需要更精确的众数估计可以先估计分布密度，再设法估计密度的最大值点。对于纯数值型的变量来说，众数作为位置特征量比较粗略，实际用处不大。

二、分散程度（变异性）特征量。不同分布之间的差别，除了大小差别（用位置特征代表），还包括分散程度的差别。比如，两个合唱小组的身高分别为 (单位：厘米)

甲组	159	160	162	160	159
乙组	150	160	180	160	150

则两个组的平均身高都是 160 厘米，但整齐程度就相差甚远。

可以衡量分散程度的特征量包括标准差、方差、变异系数、极差、四分位间距。

总体 Y 的方差为 $\mathrm{Var}(Y)=E(Y-EY)^2$, 标准差为 $\sigma_Y=\sqrt{\mathrm{Var}(Y)}$。样本 $y_1,y_2,\ldots,y_n$ 的样本方差和样本标准差分别为

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

其中分母 $n-1$ 在 $y_1,y_2,\ldots,y_n$ 是 Y 的独立重复抽样时保证了 S^2 是 $\mathrm{Var}(Y)$ 的无偏估计。在 R 语言中用 `var(x)` 求样本方差，用 `sd(x)` 求样本标准差。

数据分析中主要使用标准差而不使用方差，这是因为标准差与原来数据具有相同量纲，而方差的量纲与原来数据量纲不同。上面的两个合唱小组的身高标准差分别为 1.225 厘米和 12.25 厘米。

因为方差和标准差基于数学期望，它们会受到极端值的影响。比如，在上面所举的月薪的例子中，标准差为 9851 元，其中经理的月薪值是一个极端值，去掉这个值之后标准差为零。

如果需要递推计算方差，记 $S_k^2=\frac{1}{k-1}\sum_{i=1}^{k}(y_i-\bar{y}_k)^2$, 可用如下递推算法：

$\bar{y}_1=y_1;\ \ S_1^2=0$

for(k **in** 2:n) {

$\quad\bar{y}_k=\frac{k-1}{k}\bar{y}_{k-1}+\frac{1}{k}y_k$

$\quad S_k^2=\frac{k-2}{k-1}S_{k-1}^2+\frac{1}{k}(y_k-\bar{y}_{k-1})^2$

}

(其中 S_1^2 仅作为初值使用)

标准差是有量纲的分散程度特征，用**变异系数**

$$\mathrm{CV} = \frac{S}{\bar{y}} \times 100\%$$

可以表示变量相对于其平均值的变化情况，是一个无量纲的量。比如，100 克的标准差，如果是来自平均 500 克一袋的奶粉，变异系数达到 20%，说明此种奶粉的装袋质量很差；如果同样是 100 克的标准差但是来自平均 50 千克一袋的面粉，则变异系数为 0.2%，说明此种面粉的装袋质量很好。

样本的**极差**定义为样本的最大值减去样本的最小值，也能比较直观地反映样本值分散程度。

设 $\hat{x}_{1/4}$ 和 $\hat{x}_{3/4}$ 是从样本 $y_1, y_2, \ldots, y_n$ 中估计的分位数 $x_{1/4}$ 和 $x_{3/4}$，称 $\hat{x}_{3/4} - \hat{x}_{1/4}$ 为**四分位间距** (Inter-Quantile Range, IQR)。四分位间距是最靠近分布中心的 50% 的样本值所占的范围大小，可以反映样本分散程度，而且不受到极端值影响。

受前面传统的中位数估计方法的启发，我们可以用如下方法计算 1/4 和 3/4 分位数。把样本从小到大排序后，如果 n 为奇数，则用 $y_{(1)}, y_{(2)}, \ldots, y_{(\frac{n+1}{2})}$ 的中位数作为 $x_{1/4}$ 的估计，用 $y_{(\frac{n+1}{2})}, y_{(\frac{n+1}{2}+1)}, \ldots, y_{(n)}$ 的中位数作为 $x_{3/4}$ 的估计。如果 n 为偶数，则用 $y_{(1)}, y_{(2)}, \ldots, y_{(\frac{n}{2})}$ 的中位数作为 $x_{1/4}$ 的估计，用 $y_{(\frac{n}{2}+1)}, y_{(\frac{n}{2}+2)}, \ldots, y_{(n)}$ 的中位数作为 $x_{3/4}$ 的估计。

三、与分布形状有关的特征量。包括**偏度**w 和**峰度**k。总体 Y 的偏度和峰度分别定义为

$$w = \frac{E(Y-EY)^3}{(\mathrm{Var}(Y))^{3/2}}, \qquad k = \frac{E(Y-EY)^4}{(\mathrm{Var}(Y))^2} - 3$$

从样本 $y_1, y_2, \ldots, y_n$ 计算的偏度和峰度分别定义为 (见 SAS Institute(2010)[35], 328–329)

$$\hat{w} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{S} \right)^3$$

$$\hat{k} = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{S} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

其中 $\bar{y}$ 和 S 分别为样本平均值和样本标准差。

偏度 w 反映了分布的对称性。以连续型分布为例，设 Y 有分布密度 $p(y)$，若 $p(y)$ 关于 EY 对称，则 $w = 0$。如果 $p(y)$ 左右两个尾部不对称，右尾长而左尾短，则 $w > 0$，称这样的分布为**右偏**分布。反之，如果 $p(y)$ 的左尾长而右尾短，则 $w < 0$, 称这样的分布为**左偏**分布。

峰度 k 反映了分布的两个尾部的衰减速度情况。如果 $p(y)$ 在 $y \to \pm\infty$ 衰减速度较慢（比如 $p(y) = O(\frac{1}{x^p}), p > 1$），则称 $p(y)$ 是**重尾 (厚尾) 分布**, 这时 $k > 0$。对于正态分布 $\mathrm{N}(\mu, \sigma^2)$，$k \equiv 0$。重尾分布的样本会有比较多的极端值，即与分布中心距离很远的值。

对于多个随机变量，可以用**相关系数**来简单描述两两之间的关系。随机变量 X 和 Y 的相关系数定义为

$$\rho = \rho(X,Y) = \frac{E[(X-EX)(Y-EY)]}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$$

这是一个取值于 $[-1,1]$ 的数，当 $|\rho|$ 接近于 1 时代表 X 和 Y 有很强的线性相关关系。对于 X 和 Y 的一组样本 $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$，相关系数估计为

$$R = \frac{\sum_{i=1}^n[(x_i-\bar{x})(y_i-\bar{y})]}{\sqrt{\sum_{i=1}^n(x_i-\bar{x})^2\cdot\sum_{i=1}^n(y_i-\bar{y})^2}}$$

也满足 $-1\le R\le 1$。

注意，相关系数只考虑了两个变量之间的线性相关性。例如，$X\sim \mathrm{N}(0,1)$，$Y=X^2$ 是 X 的函数，但是 X 和 Y 的相关系数等于零。实际数据中，样本相关系数绝对值很小不表明两个变量相互独立；样本相关系数绝对值较大，则两个变量之间很可能是不独立的，但不能仅靠相关系数确认两个变量之间有线性相关性：非线性的相关也可能导致相关系数绝对值较大。

如果我们有 p 个随机变量 $(Y_1,Y_2,\ldots,Y_p)^T$(T 表示矩阵转置)，看成随机向量 $\boldsymbol{Y}$，$\boldsymbol{Y}$ 的期望为 $E\boldsymbol{Y}=(EY_1,EY_2,\ldots,EY_p)^T$，协方差阵为 $\mathrm{Var}(\boldsymbol{Y})=E[(\boldsymbol{Y}-E\boldsymbol{Y})(\boldsymbol{Y}-E\boldsymbol{Y})^T]$, $\mathrm{Var}(\boldsymbol{Y})$ 的第 k 行第 j 列元素为

$$\mathrm{Cov}(Y_k,Y_j)=E[(Y_k-EY_k)(Y_j-EY_j)],\ k,j=1,2,\ldots,p.$$

$\boldsymbol{Y}$ 的相关阵 R 是 $\boldsymbol{Y}$ 的各分量的相关系数组成的矩阵，主对角线元素等于 1，第 k 行第 j 列元素为 $\rho(Y_k,Y_j)$。

对 $\boldsymbol{Y}$ 进行 n 次独立观测得到样本 $(y_{i1},y_{i2},\ldots,y_{ip})$, $i=1,2,\ldots,n$, 用

$$\bar{y}_{\cdot j}=\frac{1}{n}\sum_{i=1}^n y_{ij},\ j=1,2,\ldots,p$$

估计 $E\boldsymbol{Y}$，用

$$\gamma_{kj}=\frac{1}{n-1}\sum_{i=1}^n(y_{ik}-\bar{y}_{\cdot k})(y_{ij}-\bar{y}_{\cdot j})$$

估计 $\mathrm{Cov}(Y_k,Y_j)$, 矩阵 $(\gamma_{kj})_{k,j=1,2,\ldots,p}$ 叫做样本协方差阵，用来估计 $\boldsymbol{Y}$ 的协方差阵。$\boldsymbol{Y}$ 的样本相关系数组成的矩阵叫做样本相关系数阵。

1.5 统计图形

图形比描述统计量更能直观、全面地反映变量分布和变量之间的关系。直方图、密度估计图、盒形图、茎叶图、QQ 图可以用来查看单个变量分布情况；散点图可以用来查看两个变量的变化关系; 三维曲面图、等值线图、数字图像可以反映自变量 x, y 与因变量 z 的关系。

1.5.1 直方图

用描述统计量可以把连续型分布的最重要的特征概括表示，但是，为了了解分布密度的形状，应该对分布密度 $p(y)$ 进行估计。比如，分布密度有两个峰的现象就是上面叙述的描述统计量无法发现的。

直方图 (histogram) 是最简单的估计分布密度的方法。

设随机变量 $Y \sim f(y)$, 分布密度 $f(y)$ 连续, $y_1, y_2, \ldots, y_n$ 为 Y 的简单随机样本。取分点 $t_0 < t_1 < \cdots < t_m$，把 Y 的取值范围分为 m 个小区间: $\cup_{k=1}^{m}(t_{k-1}, t_k]$；令

$$\begin{aligned} p_k =& P\{t_{k-1} < Y \le t_k\} \\ =& \int_{t_{k-1}}^{t_k} f(y)dy = f(\xi_k)(t_k - t_{k-1}),\ k = 1, 2, \ldots, m \end{aligned}$$

设样本值 $y_1, y_2, \ldots, y_n$ 落入第 k 个小区间 $(t_{k-1}, t_k]$ 的个数为 u_k (称为频数), 用 u_k/n 估计 p_k, 称 u_k/n 为样本落入第 k 个小区间的**频率**或百分比。当 $n \to \infty$ 时，取分点 $\{t_k\}$ 使小区间长度趋于零，即

$$d = \max\{t_k - t_{k-1} : k = 1, 2, \ldots, m\} \to 0.$$

如果 Y 的取值范围没有下界, 则要求 $t_0 \to -\infty$; 如果 Y 的取值范围没有上界, 则要求 $t_m \to +\infty$。这时有 $u_k/n \to p_k$, $k = 1, 2, \ldots, m$, a.s., 于是

$$\left| \frac{u_k}{n(t_k - t_{k-1})} - f(y) \right| \to 0,\ \text{a.s.},\ y \in (t_{k-1}, t_k],$$

即 $f(x)$ 可以用 x 所在的小区间上的频率除以小区间长度来估计。令

$$f_n(y) = \begin{cases} \sum_{k=1}^{m} \frac{u_k/n}{t_k - t_{k-1}} I_{(t_{k-1}, t_k]}(y), & \text{当 } t_0 < y \le t_m \\ 0, & \text{当 } y \le t_0 \text{ 或 } y > t_m \end{cases}$$

则 $n \to \infty$ 时 $f_n(y) \to f(y)$, a.s.。另外，$f_n(y) \ge 0, y \in (-\infty, \infty)$, $\int_{-\infty}^{\infty} f_n(y)dy = 1$, 满足密度函数的一般要求。

根据上述讨论，我们可以用如下的等距概率直方图来估计分布密度 $f(y)$。确定区间端点 (a,b) 使得样本值 $y_1,y_2,\ldots,y_n$ 都落入 (a,b) 内，确定分组数 m，令组距 $h=\frac{b-a}{m}$, 分点

$$t_k=a+kh,\ k=0,1,\ldots,m,$$

计算样本值落入第 k 个小区间的个数 u_k 和频率 u_k/n。以 $(t_{k-1},t_k]$ 为底，$(u_k/n)/(t_i-t_{i-1})=(u_k/n)/h$ 为高画长方形, 这 m 个长方形的顶部为密度估计值。

另一种等距直方图用小区间对应的频数 u_k 作为这些长方形的高度，这样的等距直方图不是密度函数的估计，但是与等距概率直方图只差常数倍 nh。

那么，如何确定分组数 m 呢？这个问题没有公认的做法。样本量 n 大时组数应当大，数据取值范围小时组数应当小，数据有效位少使得不同值少时组数应当小，每一小区间不应为空，达到 5 以上更好。实际取 m 和分点时，还需要使得分点的小数位数较少。

按照 AMISE(平方误差积分渐近期望) 最小原则，可取区间长度 h 为

$$h=3.49\sigma n^{-1/3}$$

其中总体标准差 σ 可用样本标准差代替，或者用 IQR 来估计 σ，得

$$h=2\cdot\mathrm{IQR}\cdot n^{-1/3}.$$

按绝对误差一致最小准则，可取

$$h=1.66S\left(\frac{\ln n}{n}\right)^{1/3}$$

其中 S 是样本标准差。

假设各个组的样本点数是二项式系数 $\binom{m-1}{k-1},k=1,2,\ldots,m$, 则

$$n=\sum_{k=1}^{m}\binom{m-1}{k-1}=2^{m-1}$$

即可取组数 m 为

$$m=1+\log_2 n.$$

关于 m 选取的讨论参见 Gentle(2002)[18] §9.2。

在 R 软件中，用 `hist(x, freq=FALSE)` 作估计分布密度的等距概率直方图，分组按默认规则分组 (见 Sturges(1926)[37])。

例 1.5.1. 下面的程序首先生成了 100 个标准正态分布随机数（可以理解为来自标准正态分布的样本量为 100 的一组简单随机样本）, 然后作了等距概率直方图:

```
set.seed(1); y <- rnorm(100)
hist(y, freq=FALSE)
```

结果见图1.1。

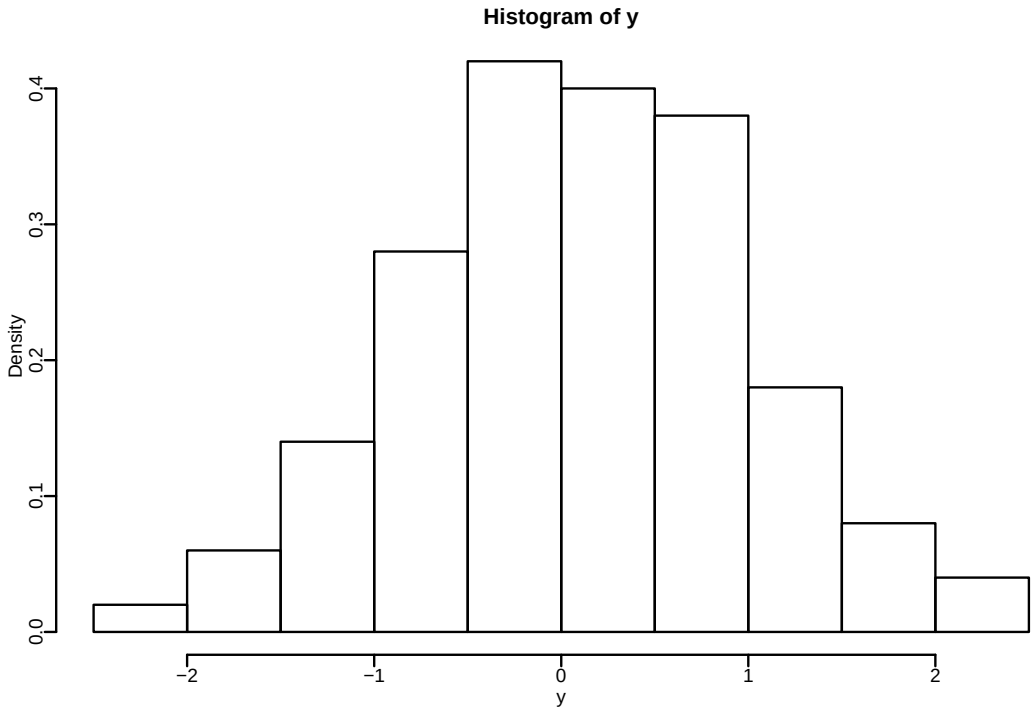

图 1.1: 100 个标准正态随机数的等距概率直方图

对于正态分布密度这样的非重尾的密度，等距概率直方图可以比较好地反映分布密度形状。但是，如果分布在两侧或一侧有重尾，这时直方图的部分小区间可能只有很少点甚至没有点，部分小区间则集中了过多的点，等距概率直方图就不能很好地反映分布密度形状。

例 1.5.2. 如下程序：

```
set.seed(3); y <- rcauchy(100)
hist(y, freq=FALSE)
```

生成了 100 个柯西分布随机数，然后对这组样本做了等距概率直方图 (见图1.2)，可以看出效果较差。

这种情况下可以把频数少的小区间合并，把频数过大的小区间拆分，产生不等距的概率直方图。比如，上面的柯西分布随机数可以用如下程序得到较好的直方图 (见图1.3)：

```
hist(y, breaks=c(-70, -30,  -10, -5, -2, -1, 0,
                 1, 2, 5, 10, 30, 70))
```

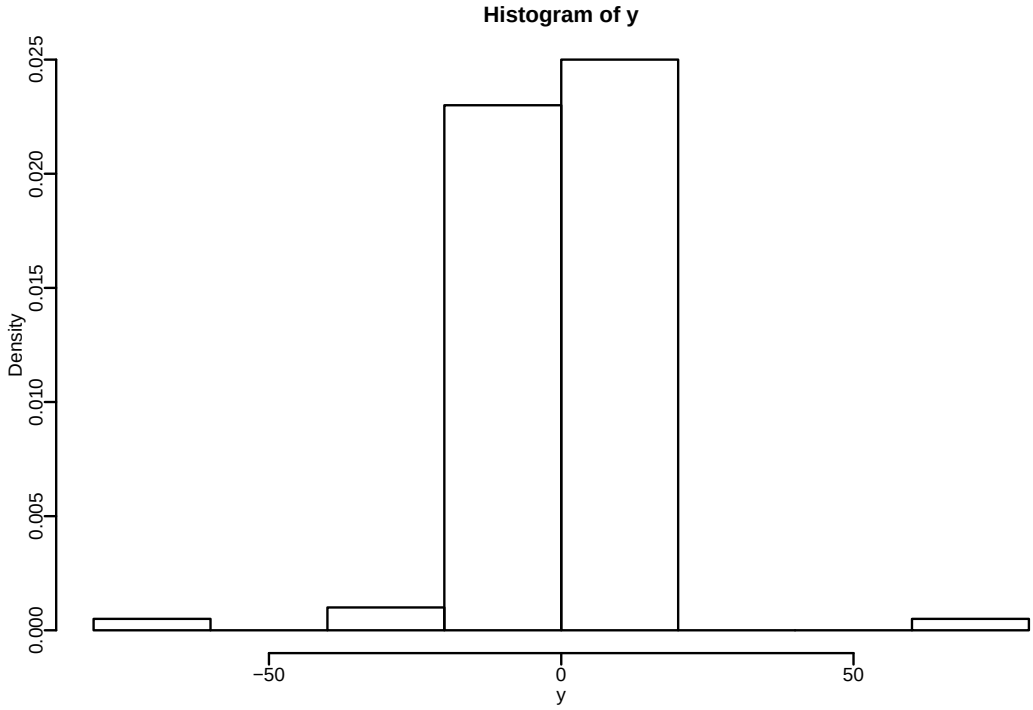

图 1.2: 100 个柯西随机数的等距概率直方图

1.5.2 核密度估计

在等距概率直方图作法中，估计 y_0 处的密度 $p(y_0)$ 时其所在区间 $(t_{k-1}, t_k]$ 的所有 $p(y)$ 都估计成同一个值，这些分点 $\{t_k\}$ 是预先取好的，与要估计 $p(y)$ 的自变量 y 的位置无关。

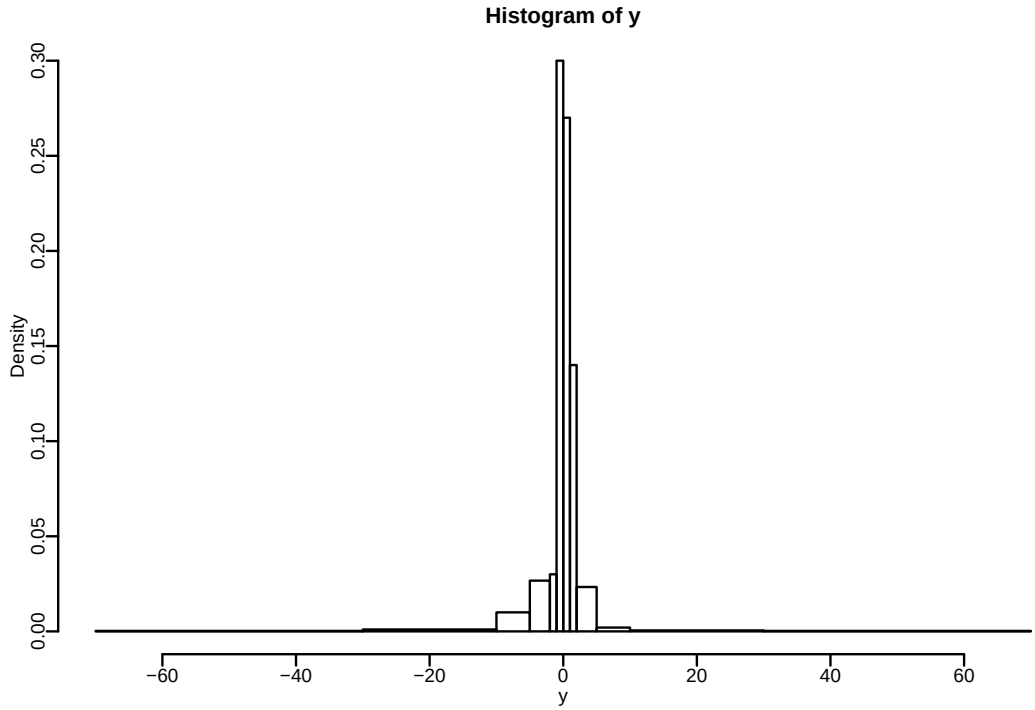

图 1.3: 100 个柯西随机数的不等距概率直方图

我们可以针对每一个 y，以 y 为中心、$\frac{h}{2}$ 为半径做小区间 $\left[y-\frac{h}{2}, y+\frac{h}{2}\right]$，用

$$\tilde{p}(y)=\frac{\#\left(\left\{y_i: y_i \in \left[y-\frac{h}{2}, y+\frac{h}{2}\right]\right\}\right)/n}{h}$$

来估计 $p(y)$，其中 $\#(A)$ 表示集合 A 的元素个数。上式可以写成

$$\begin{aligned}\tilde{p}(y) =& \frac{1}{h}\frac{1}{n}\sum_{i=1}^{n} I_{\left[y-\frac{h}{2}, y+\frac{h}{2}\right]}(y_i) \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h} I_{\left[-\frac{1}{2}, \frac{1}{2}\right]}\left(\frac{y-y_i}{h}\right),\ y \in (-\infty, \infty).\end{aligned}$$

这个分布密度估计叫做 Rosenblatt 直方图估计。如果把上面的区间改为左开右闭区间 $\left(y-\frac{h}{2}, y+\frac{h}{2}\right]$, $\tilde{p}(y)$ 与经验分布函数 $F_n(y)$ 恰好满足如下关系：

$$\tilde{p}(y)=\frac{F_n\left(y+\frac{h}{2}\right)-F_n\left(y-\frac{h}{2}\right)}{h},\ y \in (-\infty, \infty).$$

即 $\tilde{p}(y)$ 是对经验分布函数用差分近似估计 $F(x)$ 导数的结果。

设函数 $K_1(x) = I_{\left[-\frac{1}{2},\frac{1}{2}\right]}(x)$, Rosenblatt 直方图估计可以写成

$$\tilde{p}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K_1\left(\frac{y-y_i}{h}\right),$$

这样形式的密度估计叫做**核密度估计**，$K_1(x)$ 叫做核函数。

一般地，核函数应该满足如下要求:

(1) $K(x) \geq 0$;

(2) $\int_{-\infty}^{\infty} K(x)dx = 1$;

(3) $K(x)$ 是偶函数;

(4) $\int_{-\infty}^{\infty} x^2K(x)dx < +\infty$。

一般核密度估计为

$$\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{y-y_i}{h}\right)$$

$K(x)$ 的条件 (1) 和 (2) 使得 $\hat{p}(y)$ 满足 $\hat{p}(y) \geq 0, y \in (-\infty,\infty)$, $\int_{-\infty}^{\infty}\hat{p}(y)dy = 1$。当 h 很小时，密度估计很不光滑，在每个 $y_{(i)}$ 处有一个尖锐的峰而没有观测值的地方密度估计为零，估计偏差小而方差大。当 h 较大时，估计比较光滑，估计偏差大而方差小。渐近地取 $h = O(n^{-1/5})$, 核密度估计的均方误差为 $O(n^{-4/5})$, 优于直方图估计，参见 Gentle(2002)[18] §9.3。

注意，核密度估计可以很容易地推广到随机向量联合分布密度的估计。

其它核函数还有：

$$\begin{aligned}
K_2(x) &= (1-|x|)I_{[-1,1]}(x) \text{ (三角形)}\\
K_3(x) &= \frac{3}{4}(1-x^2)I_{[-1,1]}(x) \text{ (二次曲线)}\\
K_4(x) &= \frac{15}{16}(1-x^2)^2I_{[-1,1]}(x) \text{ (双二次核)}\\
K_5(x) &= \frac{70}{81}(1-|x|^3)^3I_{[-1,1]}(x) \text{ (双三次核)}\\
K_6(x) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \text{ (高斯核)}
\end{aligned}$$

R 软件中函数 `density(x)` 进行核密度估计。

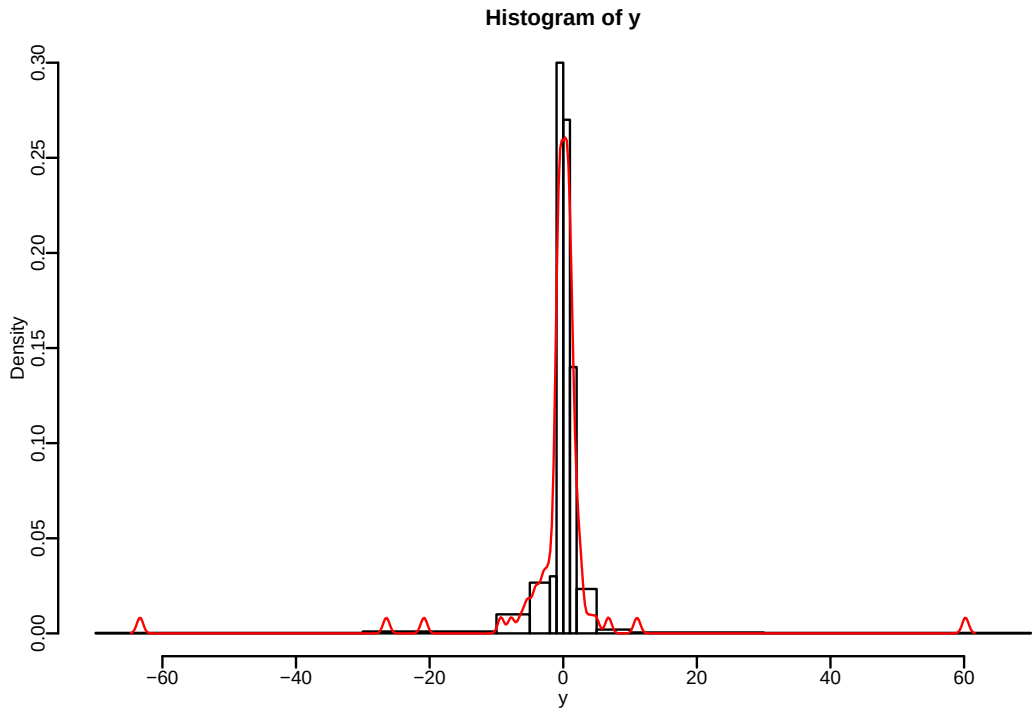

图 1.4: 100 个柯西随机数的不等距概率直方图与核密度曲线

例 1.5.3. 例1.5.2中的柯西分布随机数的直方图可以用如下程序增加核密度估计曲线：

```
res <- density(y); lines(res, col='red')
```

见图1.4。

1.5.3 盒形图

对于单峰的分布密度，**盒形图** (boxplot) 可以概括地表现其分布情况。图1.5是 100 个对数正态分布随机数的直方图与盒形图，直方图上叠加了估计的对数正态密度曲线。对数正态分布是右偏的单峰分布。盒形图中间有条粗线，位于分布的中位数上；盒子的下边缘和上边缘分别位于分布的 1/4 分位数和 3/4 分位数处，所以盒形的范围内包括了样本值分布从小到大排列中间的 1/2 的样本值。在盒子两端的外侧分别向下和向上延伸出两条线，叫做**触须线**，触须线长度小于等于 1.5 倍 IQR(即盒子高度) 并最长只能延伸到最小值或最大值的地方。如果有样本值超出了触须线的范围，就用散点符号画出（如图1.5的盒形图上端的圈），这些样本

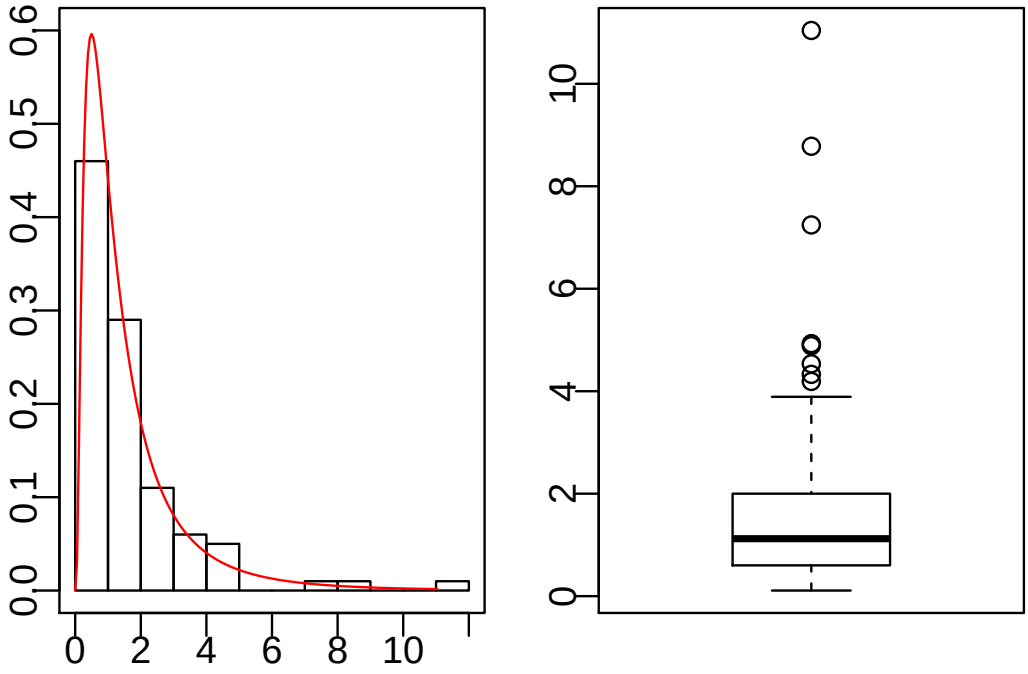

图 1.5: 100 个对数正态随机数的概率直方图与盒形图

值距离分布中心较远，称为**离群值** (outliers)。如果有较多、较远的离群值，这个样本分布就是重尾的。

所以，我们能够从盒形图上看到中位数、1/4 分位数、3/4 分位数、最小值、最大值的位置，并能发现离群值，判断分布左端和分布右端是否有重尾。另外，比较中位数上方和中位数下方的长度可以发现分布是对称的、左偏的，还是右偏的。图1.5的盒形图是右偏的。

在 R 软件中，用 `boxplot(x)` 作盒形图。

1.5.4 茎叶图

茎叶图是一种低精度的图形，用来检查数据输入和分布形状。在 R 软件中用 `stem(x)` 作茎叶图。

例 1.5.4. 如下程序输入了若干个人的身高（单位: 厘米）并作了茎叶图，

```
y <- c(145, 150, 155, 156, 159, 160, 166, 170, 190)
```

```
stem(y)
```

结果如下

```
The decimal point is 1 digit(s) to the right of the |

14 | 5
15 | 0569
16 | 06
17 | 0
18 |
19 | 0
```

图形分为竖线左边的“茎”和右边的“叶”两部分。茎部分构成了纵坐标轴，小数点位置在竖线的地方，但是根据数据范围的不同可能需要调整，比如上图的说明部分约定了小数点位置是向竖线右方移动一位（即乘以 10）。叶部分的每个数字是一片叶子，代表了一个样本点，叶子是样本值最后一个有效数字（或四舍五入后的值）。比如，茎“15”右边有 5 片叶子，分别代表样本值 150、155、156、159。

把这个图逆时针旋转 90° 来看，可以看成是一个直方图，某个茎右侧的叶子越多，分布密度在这个值附近越大。

图中的 190 明显远离了其它值。由此可见，从茎叶图可以比较容易地发现离群点。

1.5.5 正态 QQ 图和正态概率图

在许多统计模型中都要假设模型中的随机变量服从正态分布。对随机变量 Y，设 $y_1, y_2, \ldots, y_n$ 是 Y 的简单随机样本，如何判断 Y 是否服从正态分布？

可以使用假设检验的方法，零假设为 Y 服从正态分布，对立假设为 Y 不服从正态分布。这样的检验有 Shapiro-Wilk 检验等。

另外，我们也可以从分布图形来直观地判断。直方图与核密度估计图可以直接查看分布密度形状是否与正态分布相近。图1.6中的四个图形叫做**正态 QQ 图**，对应的样本分别来自为正态分布、对数正态分布、指数分布的相反数、t(2) 分布的随机数。

正态 QQ 图包括一组散点和一条直线。设样本次序统计量为 $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$，则 $y_{(i)}$ 可以作为是总体的 i/n 分位数的估计，参见1.4.1关于样本分位数的说明。设标准正态分布的分布函数为 $\Phi(\cdot)$，令 $x_i = \Phi^{-1}(i/n)$，则 $(x_i, y_{(i)})$ 分别是标准正态分布 i/n 分位数和样本的 i/n

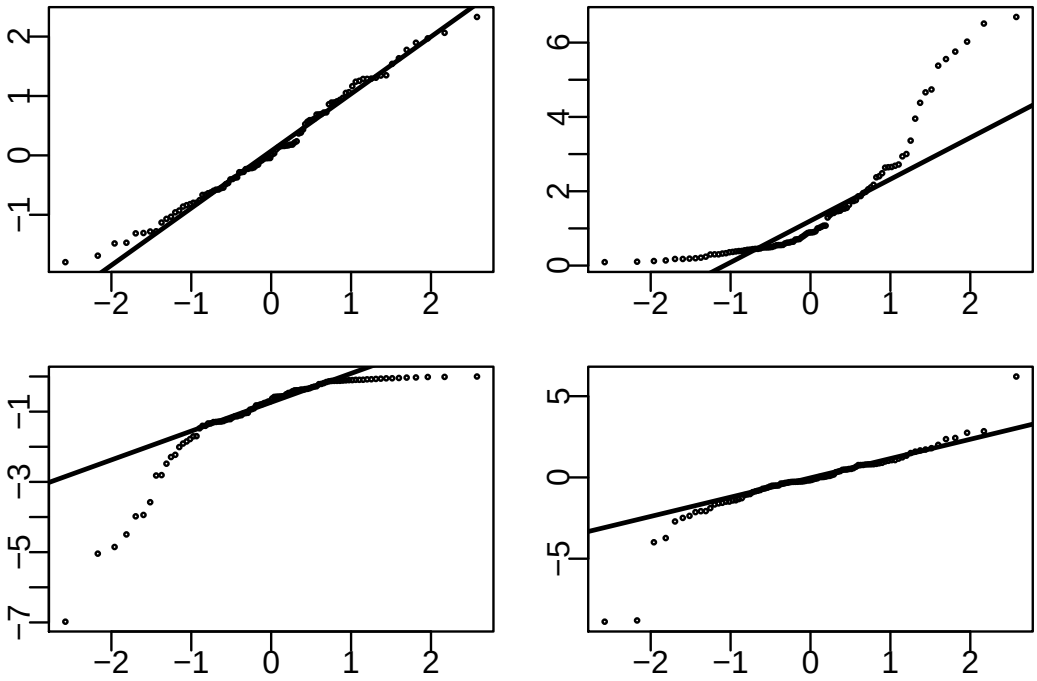

图 1.6: 正态分布、对数正态分布、指数分布取相反数、t(2) 分布的正态 QQ 图

分位数，以 $(x_i, y_{(i)}), i = 1, 2, \ldots, n$ 为坐标在直角坐标系中作图就得到了正态 QQ 图中的散点。通过代表 1/4 分位数和 3/4 分位数的两个点作直线，就是正态 QQ 图中的那些直线。

如果样本来自正态 $N(\mu, \sigma^2)$, 记 $\mathrm{N}(\mu, \sigma^2)$ 的分布函数为 $F(x)$，易见 $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$。因为 $y_{(i)}$ 是样本的 i/n 分位数，应该近似等于 $F(x)$ 的 i/n 分位数 $F^{-1}(i/n)$，所以

$$\begin{aligned} y_{(i)} \approx & F^{-1}(i/n) \\ F(y_{(i)}) \approx & \frac{i}{n} \\ \Phi\left(\frac{y_{(i)} - \mu}{\sigma}\right) \approx & \frac{i}{n} \\ \frac{y_{(i)} - \mu}{\sigma} \approx & \Phi^{-1}(i/n) = x_i \\ y_{(i)} \approx & \mu + \sigma x_i \end{aligned}$$

可见这时正态 QQ 图中的散点应该落在以 μ 为截距、以 σ 为斜率的直线附近。图1.6的第一个图就是这样的情况。

上述讨论中我们取 $y_{(i)}$ 作为总体的 i/n 分位数的估计，这样，最小值 $y_{(1)}$ 代表 $1/n$ 分位数而最大值 $y_{(n)}$ 代表 100% 分位数，两侧不对称。实际作图时，$y_{(i)}$ 对应的概率 p 可能会进行一些修正；比如，SAS 软件中的正态 QQ 图把 $y_{(i)}$ 作为 $(i-\frac{3}{8})/(n+\frac{1}{4})$ 分位数的估计。

如果样本来自于右偏的分布，正态 QQ 图就会呈现出图1.6第二个图那样的下凸形状。这是因为，图中直线上端代表了正态分布情况下在 3/4 分位数之后应有的走势，图中散点的上端高于代表正态分布的直线，说明实际数据的分布密度的右端比正态分布密度要长；此图的左端的散点高于代表正态分布的直线，说明实际数据的分布密度的左端比正态分布密度要短。这正是右偏分布的典型特征。

类似地，图1.6第三个图是典型的左偏分布的形状。而第四个图则是分布基本对称但两端的尾部都比正态分布长，从而是对称重尾分布的典型形状。

在 R 软件中，用

```
qqnorm(x); qqline(x)
```

作正态 QQ 图。

正态 QQ 图的一个变种是正态概率图，这时散点坐标和直线都不变，但是横坐标轴的刻度不是 QQ 图中 x_i 的值，而是标为 $\Phi(x_i)$ 的值，即 (x_i, y_i) 所对应的概率值。

1.5.6 散点图和曲线图

设有两个变量 X 和 Y 以及它们的 n 组观测值 $(x_i, y_i), i = 1, 2, \ldots,$。以 (x_i, y_i) 为坐标画点，把这 n 组观测值画在平面直角坐标系中，叫做**散点图**。

散点图能够比较直观地表现两个变量之间的关系，两个变量的取值范围，取值的聚集、分组，离群点等情况。

图1.7的第一个图中的 X 和 Y 有线性相关关系，且 X 增加 Y 也倾向于增加。第二个图中的 X 和 Y 有非线性相关关系，X 增加时 Y 倾向于减少。第三个图中的 X 和 Y 没有明显的相关，也没有明显的聚集模式。第四个图中的 X 和 Y 的观测值明显地分成两组，且在每一组内部 X 和 Y 的值是负相关的。注意，如果对第四组数据不画图而直接作线性回归，结果也会有显著的线性相关，但是 X 和 Y 会被误认为有正相关。

在 R 软件中用 `plot(x, y)` 画散点图。

图1.8的第一个图是 200 对观测值的散点图。图中的点是在水平方向和竖直方向对齐的。这是因为变量取值是离散的。这样的图有一个问题：很多点会重合在一起，使我们不容易发现数据的规律。R 软件的 `jitter(x, amount)` 函数可以对输入的 `x` 加一个 ±`amount` 幅度内的扰动，对扰动后的 `x` 和 `y` 作散点图就可以避免过多的点重合。如图1.8的第二个图。

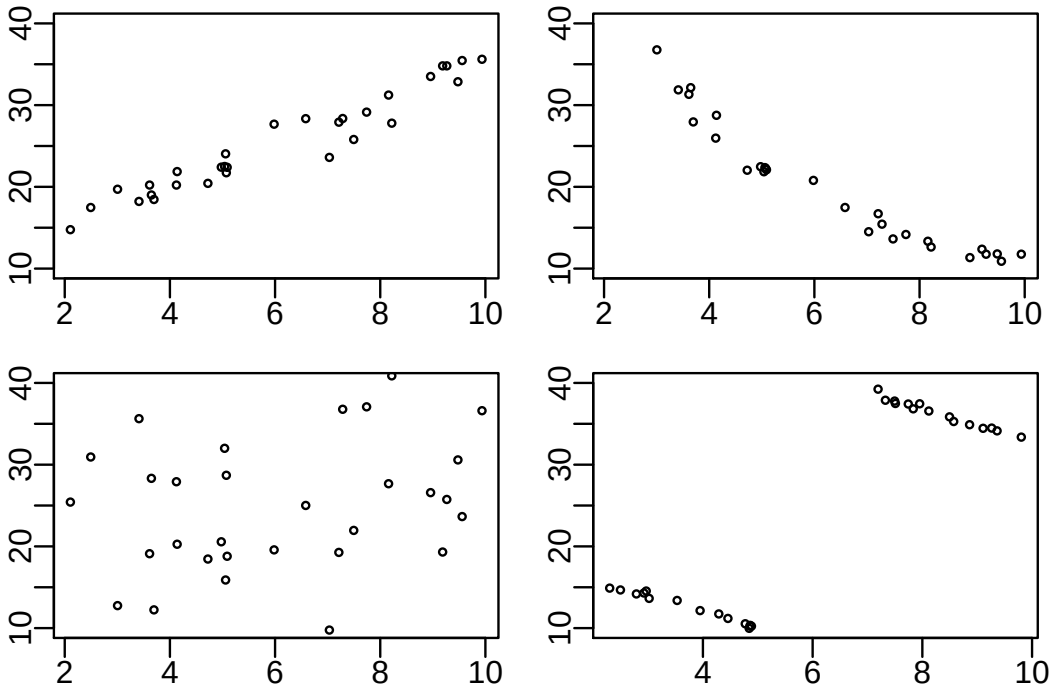

图 1.7: 散点图的例子

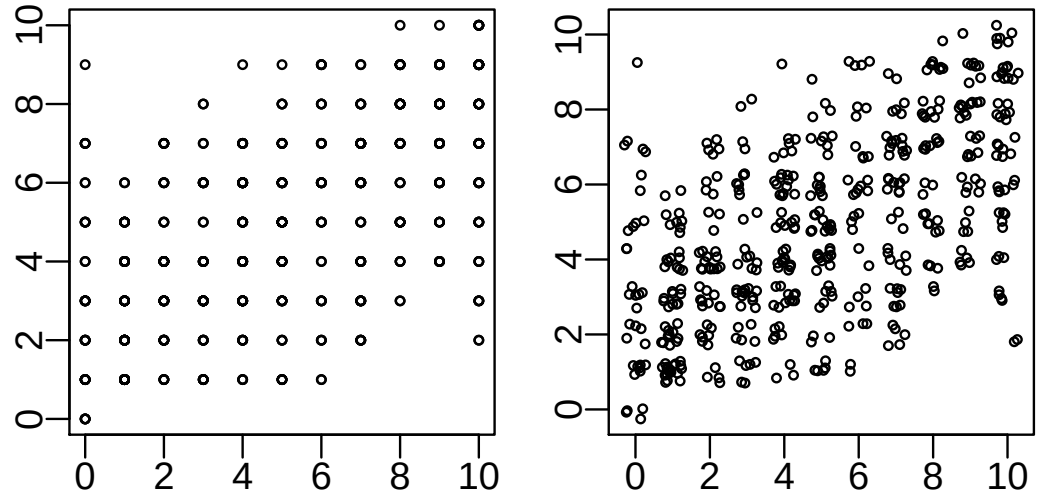

图 1.8: 离散变量散点图的例子

有些数据是随时间变化的，而数据中有时已经有时间变量，也可能没有明显的时间变量。在没有时间变量时，有时输入数据的次序号可以作为时间变量。为了考察随时间变化的量（称为时间序列），可以把以此变量为纵坐标、以时间为横坐标的点用线相连，称这样的图为**曲线图**或时间序列图，实际上，这样的图是折线图，曲线图是一种习惯称呼。

在 R 中，用 `plot(x, y, type="l")` 作曲线图。用 `lines(x, y)` 在已有图形上添加线，用 `points(x, y)` 在已有图形上添加点，用 `abline(h=...)` 在已有图形上添加横线，用 `abline(v=...)` 在已有图形上添加竖线。用 `rug` 函数可以在把点的横坐标值或纵坐标值用短线标在坐标区域边缘。

例 1.5.5. 如下 R 程序作了正弦曲线和余弦曲线图：

```
x <- seq(0, 2*pi, length=100)
y1 <- sin(x); y2 <- cos(x)
plot(x, y1, type='l',
     lwd=2, main='', ylab='')
lines(x, y2, lwd=2, lty=2)
abline(h=0, lwd=2)
abline(v=0, lwd=2)
```

其中 `lwd` 指定线的粗细，`lty` 规定线型（实线、虚线、点划线等）。见图1.9。

R 函数 `matplot` 可以在同一坐标系中做多条曲线或多组散点图。

R 中有专门的时间序列数据类型，这种数据类型是向量的变种，定义了序列的开始时间和采样频率（比如每月一次的数据的采样频率为 12），所以序列中每个数值都有对应的时间。用 `ts` 函数定义时间序列。如果 `y` 是时间序列, `plot(y)` 可以直接作时间序列曲线图。图1.10是美国泛美航空公司 1949—1960 年每月国际航班订票数的时间序列图。

对于时间序列，还可以以 $(y_i, y_{i+1}), i = 1, 2, \ldots, n-1$ 为坐标画散点图以查看自相关情况。时间序列还有一些专门的图形，如自相关函数图、偏相关函数图、谱密度估计图等。

有时随时间变化的是变量间的关系，这就更困难，可以做一系列随时间变化的图形。

散点图、曲线图等图形的坐标轴刻度一般是自动确定的，大致取为该维数据的最小值到最大值的范围。如果需要比较多幅图，要注意坐标范围的问题，必要时可以使多幅图的坐标范围统一化。

曲线图的宽高比不仅影响视觉效果，不合适的宽高比可能误导我们对变化规律的认识。在图1.11中有两幅时间序列图，实际是同一时间序列，但是下面的图清楚地表现了一个有先升后降趋势的序列，而上面的图表面上看是一个在某一水平线上下波动的序列。

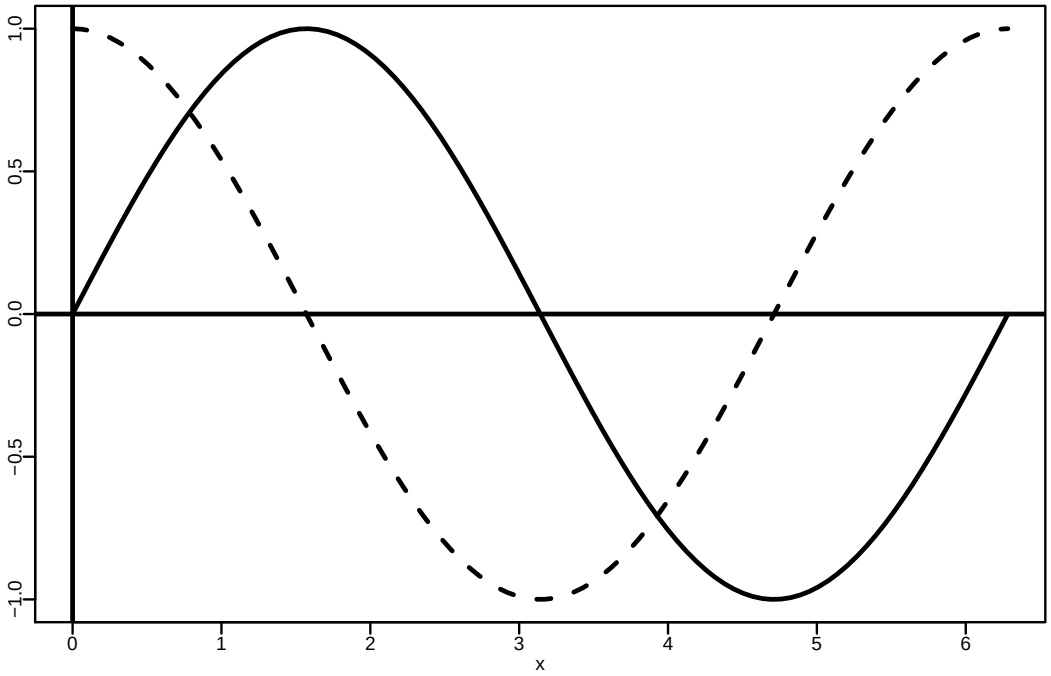

图 1.9: 曲线图

1.5.7 三维图

如果有多个数值型变量，可以两两作散点图，构成散点图矩阵。在 R 中可以用 `pairs` 函数作散点图矩阵。图1.12是三种不同的鸢尾花的 150 个样品的花瓣、花萼长、宽的数据的散点图矩阵。

对于三个变量 X, Y, Z 之间的关系，可以用各种三维图形直观地查看。

为了查看二元函数 $z = f(x, y)$ 的图像，可以作曲面图、等值线图和栅格图。R 软件的 `surface(x, y, z)` 作曲面图，其中向量 `x` 和向量 `y` 的值构成 Oxy 平面上的一张网格，`z` 为一个矩阵，第 i 行第 j 列元素为 $f(x_i, y_j)$ 的值。R 函数 `contour(x, y, z)` 作等值线图，这也是反映曲面 $f(x, y)$ 形状的图形，想法来自于地图中的等高线。R 函数 `image(x, y, z)` 作栅格图，每个 (x_i, y_j) 处用不同的颜色或灰度代表不同的 $z = f(x_i, y_j)$ 的值。

例 1.5.6. 设随机向量 (X, Y) 服从联合正态分布，$EX = EY = 0$, $\mathrm{Var}(X) = \sigma_1^2$, $\mathrm{Var}(Y) = \sigma_2^2$,

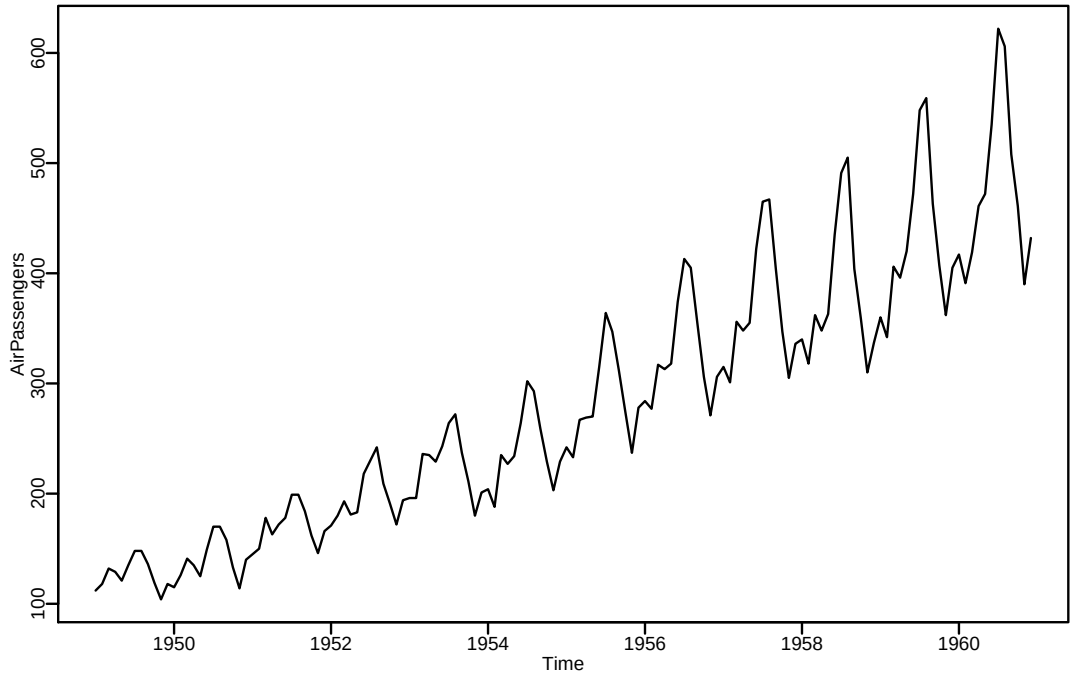

图 1.10: 1949—1950 年每月国际航班订票数的时间序列图（单位：千人）

$\rho(X,Y)=R$。则 (X,Y) 的联合密度为

$$f(x,y)=\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-R^2)}}\exp\left\{-\frac{1}{2}\frac{\sigma_2^2x^2+\sigma_1^2y^2-2R\sigma_1\sigma_2xy}{\sigma_1^2\sigma_2^2(1-R^2)}\right\}$$

如下的 R 程序作 $f(x,y)$ 的曲面图、等值线图和栅格图：

```
ng <- 30
s1 <- 1; s2 <- 2
R <- 0.5
x <- seq(-3, 3, length=ng)*s1
y <- seq(-3, 3, length=ng)*s2
d <- s1^2 * s2^2 * (1 - R^2)
f <- function(x, y) 1 / (2*pi*sqrt(d)) *
     exp(-0.5 * (s2^2 * x^2 + s1^2 * y^2 - 2*R*s1*s2*x*y)/d)
z <- outer(x, y, f)
```

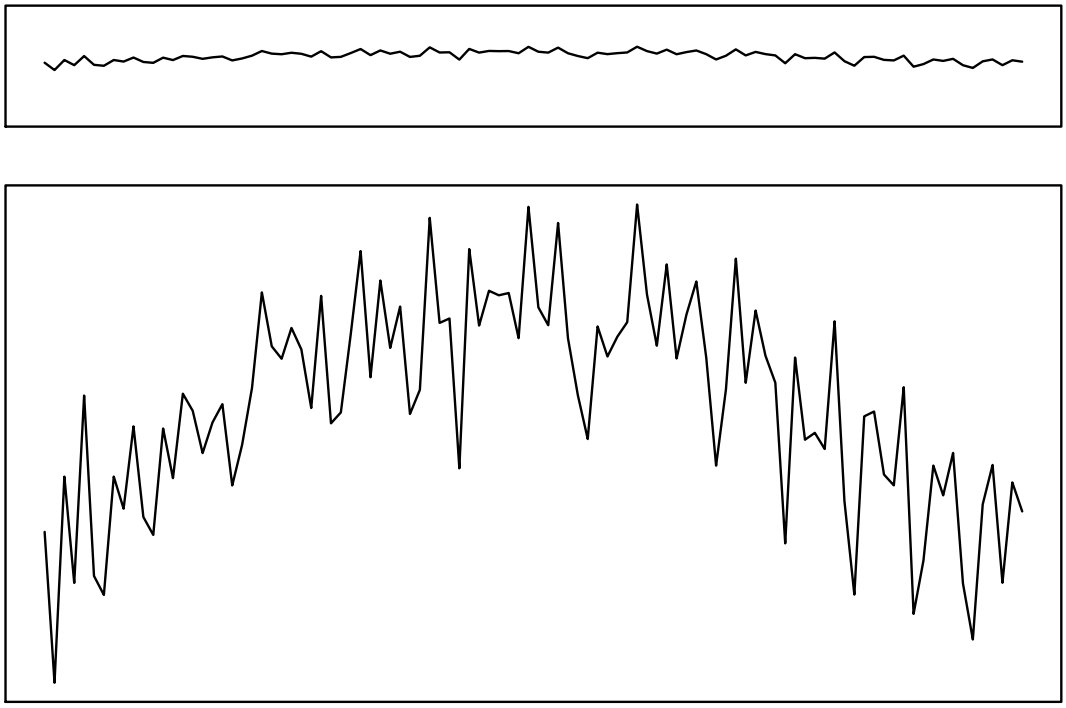

图 1.11: 同一时间序列的两幅宽高比不同的曲线图

```
persp(x, y, z, theta=-20, phi=30, r=3)
image(x, y, z, col=gray((0:32)/32))
```

见图1.13和图1.14。程序中, `outer(x, y, f)` 表示对 `x` 的每个元素 x_i 和 `y` 的每个元素 y_j 计算 $f(x_i, y_j)$，输出一个矩阵作为结果，矩阵的第 i 行第 j 列元素为 $f(x_i, y_j)$。`persp` 函数的 `theta`、`phi`、`r` 参数用来调节三维图的视角。在等值线图中，等值线稀疏的地方是函数 $f(x, y)$ 变化缓慢的地方，等值线密集的地方是函数 $f(x, y)$ 变化剧烈的地方。`image` 函数的 `col` 参数给出表示不同 `z` 值的颜色表，函数 `gray(level)` 表示灰色，level 取值于 0 到 1 之间，0 表示完全黑暗，1 表示完全明亮。栅格图中越亮的 (x, y) 坐标处的 z 值越大。

类似于散点图，三维散点图可以反映三个变量之间的关系。在 R 中，可以用 `lattice` 包的 `cloud` 函数作三维散点图。

如果变量 X 和 Y 是数值型的，变量 Z 是分类的，则可以画 (X, Y) 散点图，但每个点根据 Z 的不同值使用不同符号（圈、三角、加号、星号、点等）或不同颜色，这样可以用散点

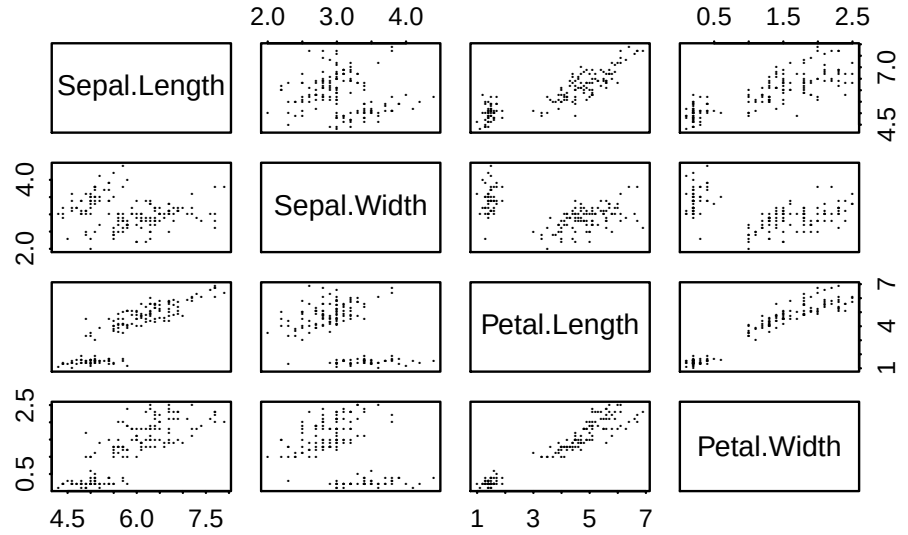

图 1.12: 鸢尾花花瓣、花萼长、宽的数据的散点图矩阵

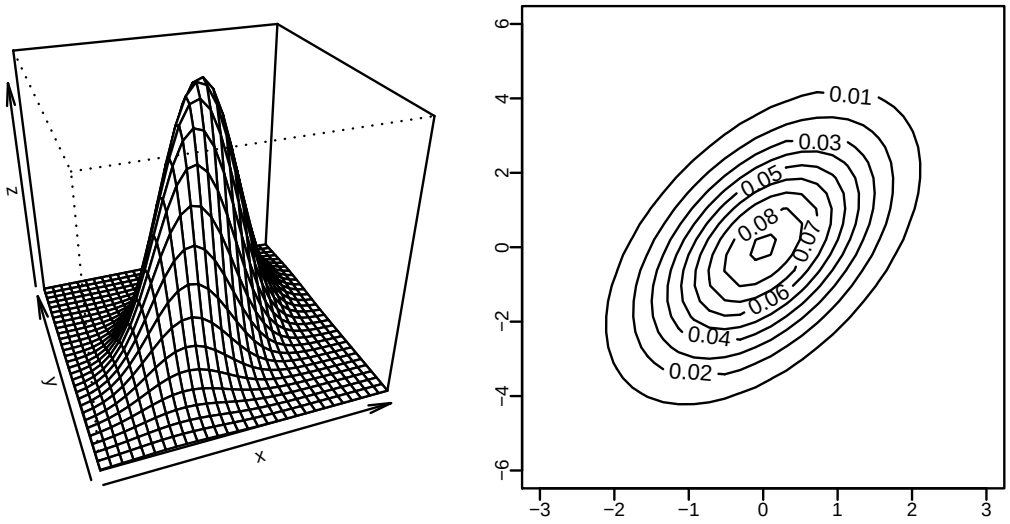

图 1.13: 二元联合正态密度曲面图和等值线图

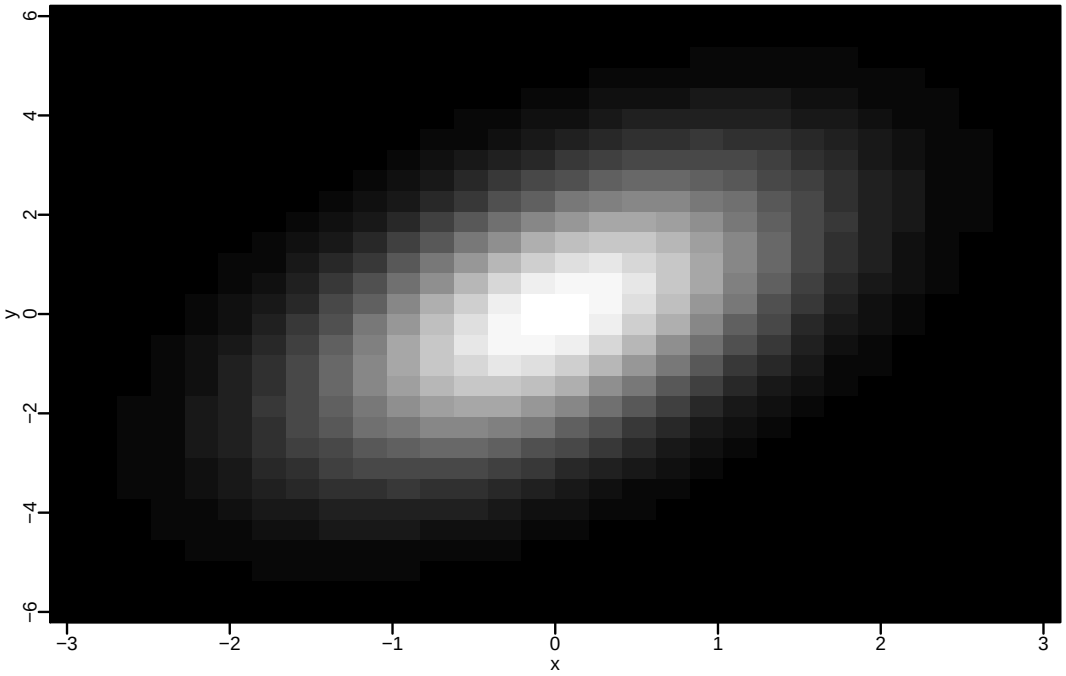

图 1.14: 二元联合正态密度栅格图

图表达出三维信息。如果变量 Z 是数值型的，可以用散点图中绘点符号的大小来表示 Z 的大小。图1.15是三种鸢尾花各 50 朵的花瓣长、宽数据的散点图，用了不同符号和颜色来代表不同种类。R 函数 `plot`、`lines`、`points` 的 `pch` 参数可以用来规定绘点符号，`col` 参数可以用来规定绘点颜色，`cex` 参数可以用来规定绘点符号的大小。

习题一

1. 编写 R 函数 `skewness(x)`, 输入向量 `x` 的值后，按如下公式

$$\phi = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3$$

   计算样本偏度，其中 n 为 `x` 的元素个数，$\bar{x}$ 和 s 分别为 `x` 中元素的样本平均值和样本标准差。

2. 设 x 是一个长度为 n 的向量, 写一段程序, 计算 x 的长度为 s 的滑动和：

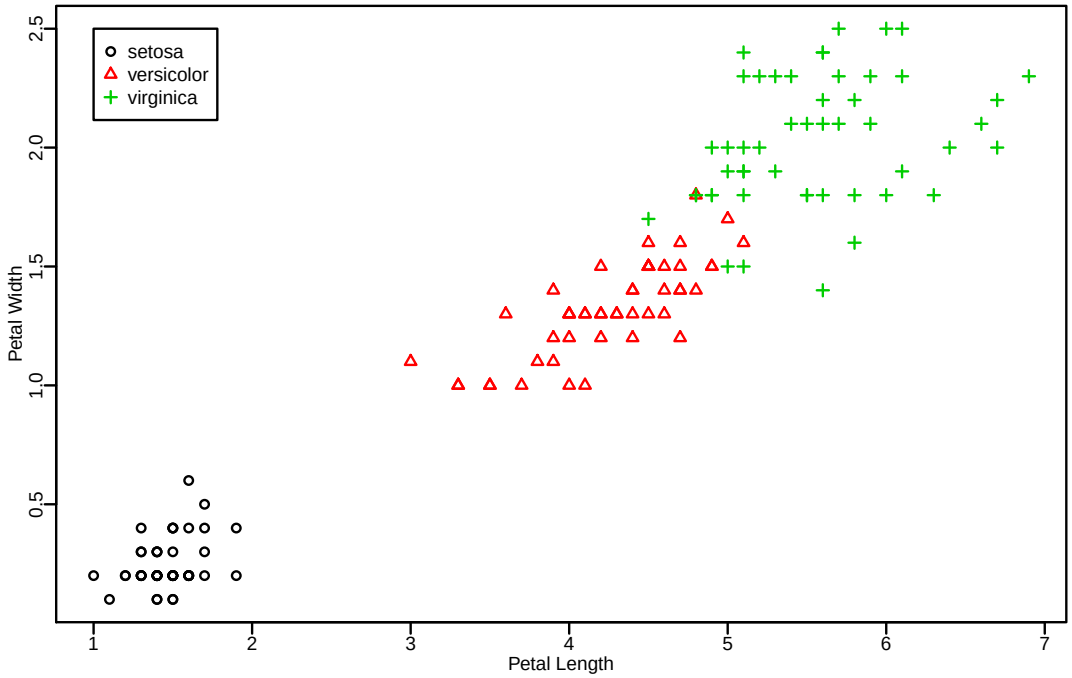

图 1.15: 三种鸢尾花的花瓣长、宽的数据

$$S_x(t) = \sum_{i=0}^{s-1} x_{t-i}, \quad t = s, s+1, \ldots, n$$

(提示：使用 `filter` 函数)

3. 写一个 AR(1) 的模拟函数：

$$x_t = a + bx_{t-1} + \varepsilon_t, t = 1, 2, \ldots, n, \mathrm{Var}(\varepsilon_t) = \sigma^2$$

函数的参数为 n、a、b、x_0 和 σ, $x_0 = x_{-1} = \cdots = x_{-p+1} = 0$，缺省时 n=100, a=0, b=1, σ=1。(提示：使用 `filter` 函数)

4. 编程计算机器 ε 的值。

5. 设某正数 a 四舍五入后保留了 p 位有效数字，表示成 $a^* = (0.a_1a_2a_3 \ldots a_p)_{10} \times 10^m$ $(a_1 \neq 0)$。估计其绝对误差和相对误差的范围，并分析误差与有效数字位数 p 的关系。

6. 如下级数

$$S_n = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots + (-1)^{n+1}\frac{1}{n}$$

收敛到 $\ln 2$。用下列四种不同方法计算 $n = 10^6$ 时级数值并比较精度：

(1) 按正常次序相加；

(2) 按相反次序相加；

(3) 成对组合后相加：$(1 - \frac{1}{2}) + (\frac{1}{3} - \frac{1}{4}) + \ldots$；

(4) 按成对组合的通分后结果相加：$\frac{1}{1\times 2} + \frac{1}{3\times 4} + \ldots$。

7. 对函数 e^{-x}，可以用两种公式计算：

$$(1)\quad e^{-x} = 1 - x + x^2/2 - x^3/3! + \cdots + (-1)^k x^k/k! + \ldots;$$
$$(2)\quad e^{-x} = 1/(1 + x + x^2/2 + x^3/3 + \cdots + x^k/k! + \ldots)$$

对 $x = 1, 2, 3, 4$，计算到 $k = 10$ 的级数并比较两种算法的计算精度。

8. 有如下一组观测数据：

$$249,\ 254,\ 243,\ 268,\ 253,\ 269,\ 287,\ 241,\ 273,\ 306$$
$$303,\ 280,\ 260,\ 256,\ 278,\ 344,\ 304,\ 283,\ 310$$

用两种方法计算样本方差：

(1) 公式

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

(2) 公式

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

其中每一个中间步骤保留 6 位有效数字 (R 中用 `signif` 函数把数据四舍五入到指定有效位数)。把结果与使用高精度计算的结果比较。

9. 设随机变量 $X \sim \mathrm{B}(30, 0.5)$, 计算 $p = P(X > 20)$，估计计算精度。用正态近似方法计算 p 并比较两个结果。

10. 对如下 t 的表达式找出发生严重精度损失的 t 值，并变化计算方法以避免精度损失：

$$(1)\frac{1}{t}-\frac{1}{t^2+a};$$
$$(2)e^{-2t^2}-e^{-8t^2};$$
$$(3)\ln(e^{t+s}-e^t);$$
$$(4)[(1+e^{-t})^2-1]t^2e^{-t}.$$

11. 编写用第26页描述的 $\tilde{F}_n$ 求逆的方法求样本分位数的程序，并用 R 的 `quantile` 函数验证。

12. 证明第28页的均值和方差的递推算法。

13. 把第28页的均值和方差的递推算法推广到随机向量的均值和协方差阵计算问题。

14. 分别对样本量 $n=10,20,30,50,100,$ 模拟生成标准正态分布、对数正态分布、柯西分布样本，并用 `hist` 函数做直方图。提示：在 R 命令行运行

```
?Distribution
```

可以获得各种与分布有关的函数说明。

15. 用 R 程序绘制图1.15, 数据为 R 中的 `iris` 数据框。

16. 设时间序列 $\{y_t\}$ 有如下模型:

$$y_t=\sum_{k=1}^{m}A_k\cos(\lambda_k t+\phi_k)+x_t,\ t=1,2,\ldots \tag{1.4}$$

其中 x_t 为线性平稳时间序列，$\lambda_k\in(0,\pi)$, $k=1,2,\ldots,m$。这样的模型称为潜周期模型。如果有 $\{y_t\}$ 的一组样本 $y_1,y_2,\ldots,y_n$, 可以定义周期图函数 (如图1.16)

$$P(\omega)=\frac{1}{2\pi n}\left|\sum_{t=1}^{n}y_te^{-it\omega}\right|^2,\ \omega\in[0,\pi]. \tag{1.5}$$

在 λ_j 的对应位置 $P(\omega)$ 会有尖峰，而且当 $n\to\infty$ 时尖峰高度趋于无穷。

如下算法可以在 n 较大时估计 m 和 $\{\lambda_k\}$：首先，对 $\omega_j=\pi j/n$, $j=1,2,\ldots,n$ 计算 $h_j\triangleq P(\omega_j)$，求 $\{h_j,j=1,2,\ldots,n\}$ 的 3/4 分位数记为 q。令 $C=qn^{1/2}$，以 C 作为分界线，设 $\{h_j\}$ 中大于 C 的下标 j 的集合为 J, 当 J 非空时，把 J 中相邻点分入一组，

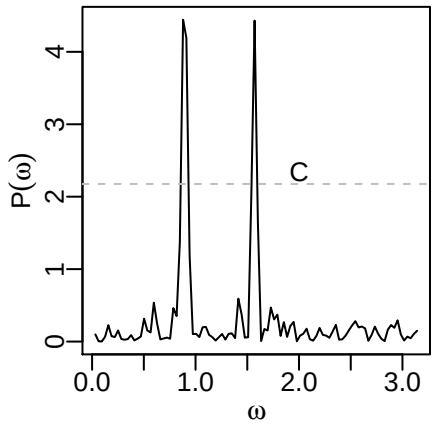

图 1.16: 潜周期模型的周期图

但是当两个下标的差大于等于 $n^{3/4}$ 时就把后一个点归入新的一组。在每组中，以该组的 h_j 的最大值点对应的角频率 $j\pi/n$ 作为潜频率 $\{\lambda_k\}$ 中的一个的估计。

用如下 R 程序可以模拟生成一组 $\{y_t\}$ 的观测数据:

```
set.seed(1); n <- 100; tt <- seq(n)
m <- 2; lam <- 2*pi/c(4, 7); A <- c(1, 1.2)
y <- A[1]*cos(lam[1]*tt) + A[2]*cos(lam[2]*tt) + rnorm(n)
```

编写 R 程序:

(1) 编写计算 $P(\omega)$ 的函数，输入 $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ 和 $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)^T$, 输出 $(P(\omega_1), P(\omega_2), \ldots, P(\omega_s))$。

(2) 用 R 函数 fft 计算 $h_j = P(\pi j/n), j = 1, 2, \ldots, n$。

(3) 对输入的时间序列样本 $y_1, y_2, \ldots, y_n$，编写函数用以上描述的算法估计 m 和 $\{\lambda_j, j = 1, 2, \ldots, m\}$。

第二章　随机数

2.1　均匀分布随机数的产生

设随机变量 X 有分布函数 $F(x)$,$\{X_i, i = 1, 2, \dots\}$ 独立同分布 $F(x)$,则 $\{X_i, i = 1, 2, \dots\}$ 的一次观测的数值 $\{x_i, i = 1, 2, \dots\}$ 叫做分布 $F(x)$ 的随机数序列，简称**随机数**。随机数是统计中一个重要的计算工具“随机模拟”的基本构成元素。

我们可以用物理方法得到一组真实的随机数，比如，反复抛掷硬币、骰子、正二十面体骰子，抽签、摇号，等等。这些方法得到的随机数质量好，但是数量不能满足随机模拟的需要。

另一种办法是预先生成大量的真实随机数存储起来，进行随机模拟时读取存储的随机数。这种方法的速度较低，已经被取代了。

现在进行随机模拟的主流方法是使用计算机实时地产生随机数，严格来说是伪随机数。**伪随机数**是用计算机算法产生的序列 $\{x_i, i = 1, 2, \dots\}$, 其表现与真实的 $F(x)$ 的独立同分布序列很难区分开来。因为计算机算法的结果是固定的，所以伪随机数不是真正的随机数；但是，好的伪随机数序列与真实随机数序列表现相同，很难区分，现在的计算机模拟都是使用伪随机数，我们把伪随机数也叫做随机数。

需要某种分布的随机数时，一般先生成均匀分布随机数，然后由均匀分布随机数再转换得到其它分布的随机数。产生均匀分布随机数的算法叫做**均匀分布随机数发生器**。

计算机中伪随机数序列是迭代生成的，即 $x_n = g(x_{n-1}, x_{n-2}, \dots, x_{n-p}), n = 1, 2, \dots, p$ 是正整数，g 是一个非线性函数。如何找到这样的 g，使得产生的序列呈现出随机性？首先要使结果有随机性。比如，把一个大的整数平方后取中间的位数，然后递推进行，叫做平方取中法。这种方法均匀性差而且很快变成零，所以已经不再使用。还可以把序列的前两项相乘然后取中间的数字，这种方法也有类似缺点。

现在常用的均匀分布随机数发生器有线性同余法、反馈位寄存器法以及随机数发生器的组合。均匀分布随机数发生器生成的是 $\{0, 1, \dots, M\}$ 或 $\{1, 2, \dots, M\}$ 上离散取值的离散均匀分布，然后除以 M 或 $M + 1$ 变成 $[0, 1]$ 内的值当作连续型均匀分布随机数，实际上是只取

了有限个值。因为取值个数有限，根据算法 $x_n = g(x_{n-1}, x_{n-2}, \ldots, x_{n-p})$ 可知序列一定在某个时间发生重复，使得序列发生重复的间隔 T 叫做随机数发生器的**周期**。好的随机数发生器可以保证 M 很大而且周期很长。

2.1.1 线性同余发生器 (LCG)

同余

定义 (同余) 设 i, j 为整数，M 为正整数，若 $j - i$ 为 M 的倍数，则称 i 与 j 关于模 M **同余**，记为 $i \equiv j \pmod{M}$。否则称 i 与 j 关于 M 不同余。

例 2.1.1.

$$\begin{aligned} 11 \equiv & 1 \pmod{10}, & 1 \equiv & 11 \pmod{10}, \\ -9 \equiv & 1 \pmod{10}. \end{aligned}$$

同余有如下性质：

(1) 对称性：$i \equiv j \pmod{M} \iff j \equiv i \pmod{M}$。

(2) 传递性：若 $i \equiv j \pmod{M}$, $j \equiv k \pmod{M}$, 则 $i \equiv k \pmod{M}$。

(3) 若 $i_1 \equiv j_1 \pmod{M}$, $i_2 \equiv j_2 \pmod{M}$, 则

$$\begin{aligned} i_1 \pm i_2 \equiv & j_1 \pm j_2 \pmod{M}, & i_1 i_2 \equiv & j_1 j_2 \pmod{M}. \end{aligned}$$

(4) 若 $ik \equiv jk \pmod{M}$(k 为正整数), 则 $i \equiv j \pmod{\frac{M}{\gcd(M,k)}}$, 其中 $\gcd(M, k)$ 表示 M 和 k 的最大公约数。

证明. (1)、(2) 和 (3) 的加减运算用同余定义很容易证明。来证明 (3) 中乘法的结果。由 $i_1 \equiv j_1 \pmod{M}$ 和 $i_2 \equiv j_2 \pmod{M}$ 的定义可知存在整数 k_1 和 k_2 使得 $j_1 - i_1 = k_1 M$, $j_2 - i_2 = k_2 M$, 于是

$$\begin{aligned} j_1 j_2 - i_1 i_2 =& j_1 j_2 - j_1 i_2 + j_1 i_2 - i_1 i_2 \\ =& j_1(j_2 - i_2) + i_2(j_1 - i_1) = j_1 \cdot k_2 M + i_2 \cdot k_1 M \\ =& (j_1 k_2 + i_2 k_1) M \end{aligned}$$

即有 $i_1 i_2 \equiv j_1 j_2 \pmod{M}$。

再来证明 (4)。设 $ik-jk=nM$, n 为整数，则 $i-j=\frac{nM}{k}$ 为整数。设 $M=M_1\cdot\gcd(M,k)$, $k=k_1\cdot\gcd(M,k)$, 则 $i-j=\frac{nM_1}{k_1}$ 且 M_1 和 k_1 互素，于是 n 被 k_1 整除，所以 $i\equiv j\pmod{M_1}$，即 $i\equiv j\pmod{\frac{M}{\gcd(M,k)}}$。 □

设 M 是正整数，A 为非负整数，A 除以 M 的余数为 $A\pmod M=A-[A/M]\times M$, 其中 $[\cdot]$ 表示取整。显然 $0\le(A\pmod M)\le M-1$，A 和 $A\pmod M$ 关于模 M 同余。在 R 中用x %% y表示 x 除以 y 的余数。

线性同余发生器

线性同余随机数发生器是利用求余运算的随机数发生器。其递推公式为

$$x_n=(ax_{n-1}+c)\pmod M,\ n=1,2,\ldots$$

这里等式右边的 $(ax_{n-1}+c)\pmod M$ 表示 $ax_{n-1}+c$ 除以 M 的余数，正整数 M 为除数，正整数 a 为乘数，非负整数 c 为增量，取某个非负整数 x_0 为初值可以向前递推，递推只需要序列中前一项，得到的序列 $\{x_n\}$ 为非负整数，$0\le x_n<M$。最后，令 $R_n=x_n/M$，则 $R_n\in[0,1)$，把 $\{R_n\}$ 作为均匀分布随机数序列。这样的算法的基本思想是因为很大的整数前面的位数是重要的有效位数而后面若干位有一定随机性。如果取 $c=0$，线性同余发生器称为**乘同余发生器**, 如果取 $c>0$，线性同余发生器称为**混合同余发生器**。

因为线性同余法的递推算法仅依赖于前一项，序列元素取值只有 M 个可能取值，所以产生的序列 $x_0,x_1,x_2,\ldots$ 一定会重复。若存在正整数 n 和 m 使得 $x_n=x_m(m<n)$，则必有 $x_{n+k}=x_{m+k},k=0,1,2,\ldots$, 即 $x_n,x_{n+1},x_{n+2},\ldots$ 重复了 $x_m,x_{m+1},x_{m+2},\ldots$，称这样的 $n-m$ 的最小值 T 为此随机数发生器在初值 x_0 下的周期。由序列取值的有限性可见 $T\le M$。

例 2.1.2. 考虑线性同余发生器

$$x_n=7x_{n-1}+7\pmod{10}.$$

取初值 $x_0=7$, 数列为: $(7,6,9,0,7,6,9,0,7,\ldots)$, 周期为 $T=4<M=10$。

例 2.1.3. 考虑线性同余发生器

$$x_n=5x_{n-1}+1\pmod{10}$$

取初值 $x_0=1$。数列为: $(1,6,1,6,1,\ldots)$, 显然周期 $T=2$。

例 2.1.4. 考虑线性同余发生器

$$x_n=5x_{n-1}+1\pmod 8$$

取初值 $x_0=1$。数列为 $(1,6,7,4,5,2,3,0,1,6,7,\ldots)$, $T=8=M$ 达最大周期。

从例子发现 $T \leq M$ 且有的线性同余发生器可以达到最大周期 M，称为满周期。满周期时，初值 x_0 取为 $0 \sim M-1$ 间的任意数都是一样的，$X_M = x_0$，序列从 X_M 开始重复。如果发生器从某个初值不是满周期的，那么它从任何初值出发都不是满周期的，不同初值有可能得到不同序列。比如随机数 $x_n = 7x_{n-1} + 7 \pmod{10}$, 从不同初值出发的序列可能为如 $7, 6, 9, 0, 7, \cdots$; $1, 4, 5, 2, 1, \cdots$; $3, 8, 3$。

不同的 M, a, c 选取方法得到不同的周期，适当选取 M, a, c 才能使得产生的随机数序列和真正的 U[0, 1] 随机数表现接近。

混合同余发生器

线性同余发生器中 $c > 0$ 时称为混合同余发生器。下列定理给出了混合同余发生器达到满周期的一个充分条件。

定理 2.1.1. 当下列三个条件都满足时，混合同余发生器可以达到满周期:

(1) c 与 M 互素;

(2) 对 M 的任一个素因子 P，$a-1$ 被 P 整除;

(3) 如果 4 是 M 的因子，则 $a-1$ 被 4 整除。

常取 $M = 2^L$，L 为计算机中整数的尾数字长。这时根据定理2.1.1的建议可取 $a = 4\alpha + 1$，$c = 2\beta + 1$，α 和 β 为任意正整数，x_0 为任意非负整数，这样的混合同余发生器是满周期的，周期为 2^L。

好的均匀分布随机数发生器应该周期足够长，统计性质符合均匀分布，序列之间独立性好。把同余法生成的数列看成随机变量序列 $\{X_n\}$，在满周期时，可认为 X_n 是从 $0 \sim M-1$ 中随机等可能选取的，即

$$P\{X_n = i\} = 1/M, \ i = 0, 1, \ldots, M-1$$

这时

$$EX_n = \sum_{i=0}^{M-1} i \cdot \frac{1}{M} = (M-1)/2,$$

$$\mathrm{Var}(X_n) = EX_n^2 - (EX_n)^2 = \sum_{i=0}^{M-1} i^2 \frac{1}{M} - \frac{(M-1)^2}{4} = \frac{1}{12}(M^2 - 1)$$

于是当 M 很大时

$$ER_n = \frac{1}{2} - \frac{1}{2M} \approx \frac{1}{2};$$
$$\mathrm{Var}(R_n) = \frac{1}{12} - \frac{1}{12M^2} \approx \frac{1}{12}$$

可见 M 充分大时从一、二阶矩看生成的数列很接近于均匀分布。

随机数序列还需要有很好的随机性。数列的排列不应该有规律，序列中的两项不应该有相关性。

因为序列由确定性公式生成，所以不可能真正独立。至少我们要求是序列自相关性弱。对于满周期的混合同余发生器可以得序列中前后两项自相关系数的近似公式

$$\rho(1) \approx \frac{1}{a} - \frac{6c}{aM}(1 - \frac{c}{M})$$

所以应该选 a 值大 (但 $a < M$)。

例 2.1.5. Kobayashi 提出了如下的满周期 2^{31} 的混合同余发生器

$$x_n = (314159269x_{n-1} + 453806245) \pmod{2^{31}}$$

其周期较长，统计性质比较好。

乘同余法 *

线性同余发生器中 $c = 0$ 时的生成方法称为乘同余法，或称积式发生器。这时的递推公式为

$$x_n = ax_{n-1} \pmod{M},\ n = 1, 2, \ldots$$

问题是如何选 M, a 达到大周期且统计性质好。显然各 x_n 都不能等于 0；如果某个 $x_k = 0$, 则 $x_{k+1} = x_{k+2} = \cdots = 0$。乘同余法有可能进入 0 就不再周期变化，称为退化情况，这时周期为 1。所以乘同余法能够达到的最大周期是 $M - 1$，每个 x_n 都只在 $\{1, 2, \ldots, M-1\}$ 中取值。

定义 (阶数) 设正整数 a 与正整数 M 互素，称满足 $a^V \equiv 1 \pmod{M}$ 的最小正整数 V 为 a 对模 M 的阶数 (或次数)，简称为 a 的阶数。

我们来证明阶数存在。考虑乘同余发生器 $x_n = ax_{n-1}$, $n = 1, 2, \ldots$, 取 $x_0 = 1$。由同余的传递性可知 $x_n \equiv a^n x_0 \pmod{M}$ 即 $x_n \equiv a^n \pmod{M}$，而 $0 \le x_n \le M - 1$ 所以 $x_n = a^n \pmod{M}$。因为 a 与 M 互素所以 $x_n \neq 0$，序列 $\{x_n\}$ 只在 $1, \ldots, M-1$ 中取值，必存在 $0 \le i < j < i + M$ 使得 $x_j = x_i$，其中 $x_i \equiv a^i \pmod{M}$，$x_j \equiv a^j \pmod{M}$，于是 $a^i \equiv a^j \pmod{M}$, 由同余的性质 (4) 有 $1 \equiv a^{j-i} \pmod{\frac{M}{\gcd(a^i, M)}}$, 其中 $\gcd(a^i, M) = 1$ 所以 $a^{j-i} \equiv 1 \pmod{M}$。取 $V = j - i$ 则 V 为正整数且 $a^V \equiv 1 \pmod{M}$。

引理 2.1.2. 设 a 与 M 互素，初值 x_0 与 M 互素，则乘同余法的周期为 a 对模 M 的阶数 V。

证明. 由同余的传递性可知 $x_V \equiv a^V x_0 \pmod{M}$, 而 $a^V \equiv 1 \pmod{M}$ 所以

$$a^V x_0 \equiv x_0 \pmod{M}$$

于是

$$x_V \equiv x_0 \pmod{M}.$$

这时必有 $x_V = x_0$，所以乘同余法的周期 $T \leq V$。

如果 $T < V$，则存在 $0 \leq i < j < i + V$ 使得 $x_j = x_i$，因 $x_j \equiv a^j x_0 \pmod{M}$, $x_i \equiv a^i x_0 \pmod{M}$, 由同余的传递性可知 $a^j x_0 \equiv a^i x_0 \pmod{M}$，由同余的性质 (4) 有

$$a^j \equiv a^i \pmod{\frac{M}{\gcd(M, x_0)}}$$

而 $\gcd(M, x_0) = 1$ 所以 $a^j \equiv a^i \pmod{M}$, 由同余的性质 (4) 可知

$$a^{j-i} \equiv 1 \pmod{\frac{M}{\gcd(a^i, M)}}$$

其中 $\gcd(a^i, M) = 1$ 所以 $a^{j-i} \equiv 1 \pmod{M}$，$0 < j - i < V$，与 V 是满足 $a^V \equiv 1 \pmod{M}$ 的最小正整数矛盾。证毕。 □

对乘同余法，当 M 与 a 互素且 M 与 x_0 互素时，$x_n = a^n x_0 \pmod{M}$，易见 x_n 与 M 互素，序列不会退化为 0.

例 2.1.6. 考虑如下乘同余发生器

$$\begin{aligned} x_n &= (8\alpha + 5) x_{n-1} \pmod{2^L} \\ x_0 &= 4b + 1 \end{aligned}$$

(其中 α, b 为非负整数)。再考虑如下的混合同余发生器

$$\begin{aligned} x_n^* &= (8\alpha + 5) x_{n-1}^* + (2\alpha + 1) \pmod{2^{L-2}} \\ x_0^* &= b \end{aligned}$$

则 $x_n = 4x_n^* + 1, n = 0, 1, 2, \ldots$。

证明. 用归纳法。当 $n=0$ 时

$$x_0^* = b,\ x_0 = 4b+1$$

所以 $x_0 = 4x_0^* + 1$ 成立。设 $x_n = 4x_n^* + 1$ 成立，由

$$x_n = (8\alpha+5)x_n \pmod{2^L},$$

其中

$$\begin{aligned}(8\alpha+5)x_n =&(8\alpha+5)(4x_n^*+1)\\ =&4[(8\alpha+5)x_n^* + (2\alpha+1)] + 1\end{aligned}$$

于是

$$\begin{aligned}(8\alpha+5)x_n \pmod{M} =&4\{[(8\alpha+5)x_n^* + (2\alpha+1)] \pmod{2^{L-2}}\} + 1\\ =&4x_{n+1}^* + 1\end{aligned}$$

得证。 □

这个例子中，由定理2.1.1可知 $\{x_n^*\}$ 是满周期的, $\{x_n\}$ 与 $\{x_n^*\}$ 一一对应，所以 $\{x_n\}$ 周期为 2^{L-2}。这里的乘同余发生器与一个混合同余发生器等价。

对乘同余法得到的随机序列 $\{X_n\}$，当 M 充分大时可以类似得到

$$ER_n \approx \frac{1}{2}, \qquad \mathrm{Var}(R_n) \approx \frac{1}{12}.$$

为了使得前后两项的相关系数较小，应取 a 较大且使得 a 的二进制表示排列无规律。

素数模乘同余法 *

若取 M 为小于 2^L 的最大素数，选取适当的 a 可以达到 $T = M-1$ 周期，这样的发生器叫**素数模乘同余发生器**。

定义 (素元) 设正整数 a 和素数 M 互素，若 a 对模 M 的阶数 V 满足 $V = M-1$，则称 a 为 M 的素元 (或原根)。

乘同余发生器中取 a 为 M 的素元可以达到最大周期 $M-1$。

例 2.1.7. $M=7$ 是素数，取 $a=3$ 与 M 互素。考虑素数模乘同余发生器

$$x_n = a^n \pmod{M},\ n = 1,2,\ldots,\ x_0 = 1,$$

序列为 $1,3,2,6,4,5,1,3,\ldots$。周期达到 $M-1=6$，所以 $a=3$ 是 $M=7$ 的素元。

可以验证 $a=5$ 也是 $M=7$ 的素元，序列为 $1,5,4,6,2,3,1,5,\ldots$。$1,2,4,6$ 不是 $M=7$ 的素元。此例说明素元可以不唯一。

在选取素数模乘同余发生器的参数 M 和 a 的时候，可以取 M 为小于 2^L 的最大素数 (L 为计算机整数表示的尾数位数)，取 a 为 M 的素元，这样保证周期 $T=M-1$。a 应尽可能大，a 的二进制表示尽可能无规律。这样在一个周期内可以得到 $1,2,\ldots,M-1$ 的一个排列。初值 x_0 可以取 $1,2,\ldots,M-1$ 中任一个。

例 2.1.8. Lewis-Goodman-Miller(1969) 的素数模乘同余发生器为

$$x_n = ax_{n-1} \pmod{2^{31}-1},\ n=1,2,\ldots,\ x_0 \text{ 为任意正整数}$$

a 取如下四个值之一: $(7^5 = 16807, 397204094, 764261123, 630360016)$。

注意，在设计线性同余法程序时，如果使用程序语言中的整数型或无符号整数型来存储 x_n，则在计算 ax_{n-1} 的乘法时可能发生溢出。对 $M=2^L$ 情形，如果程序语言支持溢出而不作为错误，则溢出恰好完成了求除以 2^L 的余数的运算。对于 $M<2^L$ 的情形也可以巧妙设计以利用溢出。对于较小的 a 和 M(如 a 和 M 都小于 $2^{L/2}$ 时) 不会溢出。如果使用双精度实数型来保存 x_n 则不需要考虑溢出问题。

例 2.1.9. 设计素数模乘同余发生器算法, 要求不产生溢出。设 a 为正整数，M 为素数，令 $b=[M/a]$, $c=M \pmod{a}$，则 $M=ab+c$, 设 $a<b$。乘同余递推中的除法满足

$$\frac{ax_{n-1}}{M} = \frac{ax_{n-1}}{ab+c} \le \frac{x_{n-1}}{b}$$

且

$$\begin{aligned}\frac{x_{n-1}}{b} - \frac{ax_{n-1}}{ab+c} &< \frac{ax_{n-1}}{ab} - \frac{ax_{n-1}}{ab+a} = \frac{x_{n-1}}{b(b+1)} \\ < \frac{M}{b(b+1)} &= \frac{ab+c}{b(b+1)} < \frac{ab+a}{b(b+1)} = \frac{a}{b} < 1\end{aligned}$$

即

$$\frac{ax_{n-1}}{M} \le \frac{x_{n-1}}{b} < \frac{ax_{n-1}}{M} + 1$$

记 $k_0 = \left[\frac{ax_{n-1}}{M}\right]$, $k_1 = \left[\frac{x_{n-1}}{b}\right]$, 则 $k_0 \le k_1 \le k_0+1$, 计算 k_0 不用计算乘法所以不会溢出，

$k_1 = k_0$ 或 $k_0 + 1$。于是

$$\begin{aligned} x_n =& ax_{n-1} - k_0 M = ax_{n-1} - k_1 M + (k_1 - k_0)M \\ =& ax_{n-1} - k_1(ab + c) + (k_1 - k_0)M \\ =& a(x_{n-1} - k_1 b) - k_1 c + (k_1 - k_0)M \\ =& a(x_{n-1} \pmod b) - k_1 c + (k_1 - k_0)M \end{aligned}$$

其中 $k_1 - k_0$ 等于 0 或 1，因为 $x_n > 0$，所以如果 $a(x_{n-1} \pmod b) - k_1 c < 0$ 则 $k_1 - k_0 = 1$。

算法如下：

设置 M, a, b, c, x_0 的值
for(n **in** $1 : N$) {
  $k_1 \leftarrow \text{floor}(x_{n-1}/b)$
  $x_n \leftarrow a(x_{n-1} - k_1 b) - k_1 c$
  if $(x_n < 0)$ $x_n \leftarrow x_n + M$
}

对例2.1.8的素数模发生器，取 $a = 16807$ 可以按上述算法计算。

上述的算法使用了伪代码来描述，伪代码的控制结构仿照 R 语言的控制结构。

2.1.2 FSR 发生器 *

线性同余法的周期不可能超过 2^L(L 为整数型尾数的位数)，而且作为多维随机数相关性大，分布不均匀。基于 Tausworthe(1965) 文章的 FSR 方法是一种全新的做法，对这些方面有改善。

FSR(反馈位移寄存器法) 按照某种递推法则生成一系列二进制数 $\alpha_1, \alpha_2, \ldots, \alpha_k, \ldots$，其中 α_k 由前面的若干个 $\{\alpha_i\}$ 线性组合并求除以 2 的余数产生:

$$\alpha_k = (c_p \alpha_{k-p} + c_{p-1}\alpha_{k-p+1} + \cdots + c_1 \alpha_{k-1}) \pmod 2,\ k = 1, 2, \ldots \tag{2.1}$$

线性组合系数 $\{c_i\}$ 只取 0, 1，这样的递推可以利用程序语言中的整数二进制运算快速实现。给定初值 $(\alpha_{-p+1}, \alpha_{-p+2}, \ldots, \alpha_0)$ 向前递推，得到 $\{\alpha_k, k = 1, 2, \ldots\}$ 序列后依次截取长度为 L 的二进制位组合成整数 x_n，$R_n = x_n/2^L$。不同的组合系数和初值选择可以得到不同的随机数发生器, 巧妙设计可以得到很长的周期，作为多维均匀随机数性质较好。

FSR 算法中系数 $(c_1, c_2, \ldots, c_p)$ 如果仅有两个为 1，比如 $c_p = c_{p-q} = 1(1 < q < p)$, 则算法变成

$$\begin{aligned}\alpha_k =&(\alpha_{k-p} + \alpha_{k-p+q}) \pmod 2\\ =&\begin{cases}0 & 若\alpha_{k-p} = \alpha_{k-p+q}\\ 1 & 若\alpha_{k-p} \neq \alpha_{k-p+q}\end{cases}\end{aligned}$$

用 $\oplus$ 表示二进制异或运算，则

$$\alpha_k = \alpha_{k-p} \oplus \alpha_{k-p+q},\ k = 1, 2, \ldots.$$

比如，取 $p = 98, q = 27$。

设计 FSR 计算机程序时，直接对包含 M 个二进制位的非负整数 $\{x_n\}$ 的数列用异或递推更方便。递推公式为

$$\begin{aligned}x_n =&x_{n-p} \oplus x_{n-p+q}, \quad (1 < q < p),\ n = 1, 2, \ldots\\ R_n =&x_n/2^M\end{aligned}$$

这需要由 p 个 M 位二进制非负整数作为种子 (初值)。这种算法只需要异或运算，不受计算机字长限制，适当选取 p, q 后周期可以达到 $2^p - 1$，作为多维随机数的性质可以很好，需要预先研究得到种子表而不能随便取初值。

2.1.3 组合发生器法

随机数设计中比较困难的是独立性和多维的分布。可以考虑把两个或若干个发生器组合利用, 可以比单个发生器由更长的周期和更好的随机性。

例 2.1.10. Wichman 和 Hill(1982) 设计了如下的线性组合发生器。

利用三个 16 位运算的素数模乘同余发生器:

$$\begin{aligned}U_n &= 171U_{n-1} \pmod{30269}\\ V_n &= 172V_{n-1} \pmod{30307}\\ W_n &= 170W_{n-1} \pmod{30323}\end{aligned}$$

作线性组合并求余:

$$R_n = (U_n/30269 + V_n/30307 + W_n/30323) \pmod 1$$

这个组合发生器的周期约有 7×10^{12} 长, 超过 $2^{32} \approx 4 \times 10^9$。

例 2.1.11. MacLaren 和 Marsaglia(1965) 设计了组合同余法，组合两个同余发生器，一个用来“搅乱”次序。

设有两个同余发生器 A 和 B。用 A 产生 m 个随机数 (如 $m = 128$)，存放在数组 $T = (t_1, t_2, \ldots, t_m)$. 需要产生 x_n 时，从 B 中生成一个随机下标 $j \in \{1, 2, \ldots, m\}$，取 $x_n = t_j$，但从 A 再生成一个新随机数 y 代替 T 中的 t_j, 如此重复。

这样组合可以增强随机性，加大周期 (可超过 2^L)。也可以只使用一个发生器，用 x_{n-1} 来选择下标。

在 R 软件中，用 `runif(n)` 产生 n 个 U(0,1) 均匀分布的随机数。R 软件提供了若干种随机数发生器，可以用 `RNGkind` 函数切换。在使用随机数进行随机模拟研究时，如果希望模拟研究的结果可重复，就需要在模拟开始时设置固定的随机数种子。虽然不同的随机数发生器种子的形式有所不同，在 R 中，总是可以用函数 `set.seed`(seed) 来设置种子，其中 seed 是一个序号，实际的种子由这个序号决定。

2.1.4 随机数的检验 *

文献中已经有许多随机数生成算法，统计软件中也已经包含了许多公认的可靠的随机数发生器。但是，我们要解决一个自己的模拟问题时，还是要反复确认所用的随机数在自己的问题中是好的，没有明显缺陷的。比如，一个经典的称为 RANDU 的随机数发生器用在一维积分时效果很好，但连续三个一组作为三维均匀分布则很不均匀。

为了验证随机数的效果，最好是找一个和自己的问题很接近但是有已知答案的问题，用随机模拟计算并考察其精度。

对均匀分布随机数发生器产生的序列 $\{R_n, n = 1, 2, \ldots, N\}$ 可以进行各种各样的检验以确认其均匀性与独立性。下面列举一些检验的想法：

- 对随机数列 $\{R_n, n = 1, 2, \ldots, N\}$ 计算

$$\bar{R} = \frac{1}{N}\sum_{n=1}^{N} R_n,\ \overline{R^2} = \frac{1}{N}\sum_{n=1}^{N} R_n^2,\ S^2 = \frac{1}{N}\sum_{n=1}^{N}(R_n - \frac{1}{2})^2$$

  则 $N \to \infty$ 时 $\bar{R}$、$\overline{R^2}$ 和 S^2 均渐近服从正态分布，可以用 Z 检验法检验这三个统计量与理论期望值的偏离程度。

- 把 $[0, 1]$ 等分成 k 段，用拟合优度卡方检验法检验 $\{R_n, n = 1, 2, \ldots, N\}$ 落在每一段的取值概率是否近似为 $1/k$。

- 用 Kolmogorov-Smirnov 检验法进行拟合优度检验，看 $\{R_n, n = 1, 2, \ldots, N\}$ 是否与 U[0,1] 分布相符。

- 把 $\{R_n, n = 1, 2, \ldots, N\}$ 每 d 个组合在一起成为 $\mathbb{R}^d$ 向量，把超立方体 $[0, 1]^d$ 每一维均分为 k 份，得到 k^d 个子集，用卡方检验法检验组合得到的 $\mathbb{R}^d$ 向量落在每个子集的概率是否近似为 k^{-d}。

- 把 $\{R_n, n = 1, 2, \ldots, N\}$ 看作时间序列样本，计算其样本自相关函数列 $\{\hat{\rho}_j, j = 1, 2, \ldots\}$，在 $\{R_n, n = 1, 2, \ldots\}$ 独立同分布情况下 $\{\sqrt{N}\hat{\rho}_j, j = 1, 2, \ldots, N\}$ 应该渐近服从独立的标准正态分布，可以据此进行白噪声检验。

- 把 $\{R_n\}$ 离散化为 $y_n = \text{floor}(kR_n), n = 1, 2, \ldots, N$, 令 $\xi_n = y_n$, $\eta_n = y_{n+b}(b$ 为正整数), $n = 1, 2, \ldots, N - b$, 用列联表检验法检验 ξ_n 和 η_n 的独立性。

- 游程检验。把序列 $\{R_n, n = 1, 2, \ldots, N\}$ 按顺序切分为不同段，每一段中的数值都是递增的，这样的一段叫做一个**上升游程**，如 (0.855), $(0.108, 0.226)$, $(0.032, 0.123)$, $(0.055, 0.545, 0.642, 0.870)$, $\ldots$。在相邻的两个上升游程中前一个游程的最后一个值大于后一个游程的第一个值。在 $\{R_n, n = 1, 2, \ldots, \}$ 独立同分布条件下游程长度的理论分布可以得知，然后可以比较实际游程长度与独立同分布条件下期望长度。

- 扑克检验。把 $\{R_n\}$ 离散化为 $y_n = \text{floor}(8R_n), n = 1, 2, \ldots, N$, 然后每连续的 8 个作为一组，计数每组内不同数字的个数（$1 \sim 8$ 个）。在 $\{R_n\}$ 独立同均匀分布条件下每组内不同数字个数的理论分布概率可以计算出来，然后用卡方检验法检验实际观测与理论概率是否相符。

- 配套检验。$\{R_n\}$ 离散化为 $y_n = \text{floor}(kR_n)(k$ 为正整数), $n = 1, 2, \ldots, N$, 然后顺序抽取 $\{y_n, n = 1, 2, \ldots, N\}$ 直到 $\{y_n\}$ 的可能取值 $\{0, 1, \ldots, k - 1\}$ 都出现过为止，记录需要抽取的 $\{y_n\}$ 的个数 L，反复抽取并记录配齐数字需要抽取的值的个数 $l_j, j = 1, 2, \ldots$。在 $\{R_n\}$ 独立同 U(0,1) 分布条件下这样的 L 分布可以得到，可以计算 $\{l_j\}$ 的平均值并用渐近正态分布检验观测均值与理论均值的差异大小，或直接用卡方检验法比较 $\{l_j\}$ 的样本频数与理论期望值。

- 正负连检验。令 $y_n = R_n - \frac{1}{2}$，把连续的 $\{y_n\}$ 的正值分为一段，把连续的 $\{y_n\}$ 的负值分为一段，每段叫做一个“连”，连长 L 的分布概率为 $P(L = k) = 2^{-k}, k = 1, 2, \ldots$，可以用卡方检验法检验 L 的分布；总连数 T 满足 $ET = \frac{n+1}{2}$, $\text{Var}(T) = \frac{n-1}{4}$, 可以用 Z 检验法检验 T 的值与理论期望的差距。

- 升降连检验。计算 $y_n = R_n - R_{n-1}$, $n = 2, 3, \ldots, N$，把连续的正值的 $\{y_n\}$ 分为一段叫做一个上升连，把连续的负值的 $\{y_n\}$ 分为一段叫做一个下降连，可以用卡方检验法比

较连的长度与 $\{R_n\}$ 独立同分布假设下的理论分布，或用 Z 检验法比较总连数与理论期望值的差距。

2.2 非均匀分布随机数的产生

设 $\{U_n, n = 1, 2, \dots\}$ 为标准均匀分布 U(0,1) 的 (伪) 随机数序列。对于其它分布 $F(x)$，我们需要产生分布为 $F(x)$、相互独立的随机数序列。产生服从某一分布 $F(x)$ 的相互独立的随机数序列也叫做从 $F(x)$ **抽样**。

2.2.1 逆变换法

定理 2.2.1. 设 X 为连续型随机变量，取值于区间 (a, b)(可包括 $\pm\infty$ 和端点), X 的密度在 (a, b) 上取正值，X 的分布函数为 $F(x)$，$U \sim \mathrm{U}(0, 1)$，则 $Y = F^{-1}(U) \sim F(\cdot)$。

证明.

$$P(Y \le y) = P(F^{-1}(U) \le y) = P(U \le F(y)) = F(y)$$

得证。 □

定理 2.2.2. 设 X 为离散型随机变量, 取值于集合 $\{a_1, a_2, \dots\}(a_1 < a_2 < \dots)$，$F(x)$ 为 X 的分布函数，$U \sim \mathrm{U}(0, 1)$，根据 U 的值定义随机变量 Y 为

$$Y = a_i \text{ 当且仅当 } F(a_{i-1}) < U \le F(a_i),\ i = 1, 2, \dots$$

(定义 $F(a_0) = 0$) 则 $Y \sim F(y)$。

证明.

$$P(Y = a_i) = P(F(a_{i-1}) < U \le F(a_i)) = F(a_i) - F(a_{i-1}) = p_i,\ i = 1, 2, \dots$$

得证。 □

用这两个定理生成随机数的方法叫做逆变换法。

2.2.2 离散型随机数

如果随机变量 X 在 $\{1,2,\ldots,m\}$ 中取值且 $P(X=i)=\frac{1}{m}, i=1,2,\ldots,m$，则称 X 服从**离散均匀分布**。

例 2.2.1 (离散均匀分布随机数). 为了生成取值于 $\{1,2,\ldots,m\}$ 的离散均匀分布样本，只要令

$$x_n=\text{ceil}(mR_n),\ n=1,2,\ldots$$

则 $\{x_n\}$ 为 m 个值的离散均匀分布随机数。函数 $\text{ceil}(x)=\min\{k:k\geq x, k\text{为整数}\}$。

例 2.2.2 (仅取有限个值的离散型随机数). 设 X 为一个仅在 $\{a_1,a_2,\ldots,a_m\}$ 中取值的离散随机变量，$P(X=a_i)=p_i, i=1,2,\ldots,m$。为了由 $U\sim \text{U}[0,1]$ 生成 X 的随机数，可以利用定理2.2.2，即当且仅当 $F(a_{i-1})<U\leq F(a_i)$ 时取 $X=a_i$。此算法依次判断 U 是否小于等于 $F(a_1)$，是否小于等于 $F(a_2)$，...，一旦条件成立就不再继续判断，所以排在前面的判断成立概率越大，平均需要的判断次数越少。只要重排 X 的取值次序使得取值概率大的排在前面就可以改进效率。改进后的算法为:

{重排 X 的取值集合为 $(b_1,b_2,\ldots,b_m)$,
对应的概率为 $(q_1,q_2,\ldots,q_m)$, 使得 $q_1\geq q_2\geq\cdots\geq q_m$ }
$F_0\leftarrow 0$
for(j **in** $1:m$) {
  $F_j\leftarrow F_{j-1}+q_j$
  }
for(n **in** $1:N$) {
  $X_n\leftarrow b_m$
  for(j **in** $1:(m-1)$) {
    if $(U_n\leq F_j)$ {
      $X_n\leftarrow b_j$
      break
    }
  }
}

注:R 中用 `order(x)` 可以获得把 `x` 的元素从小到大排列的下标,比如,`order(c(3,1,7,4))` 结果为 $(2,1,4,3)$,`x[order(x)]` 是 `x` 的元素从小到大排列结果,用 `order(x, decreasing=TRUE)` 可以获得把 `x` 的元素从大到小排列的下标。R 中用 `sample(x, size=n, prob=p,`

`replace=TRUE`) 可以生成样本量为 `n` 的有限个值的离散型随机变量随机数，其中向量 `x` 是变量的可取值集合, 向量 `p` 是每个值对应的概率。

例 2.2.3 (随机排列和无放回随机抽样). 生成 $(1,2,\ldots,n)$ 的一个随机排列。第一种想法是从 $A=(1,2,\ldots,n)$ 中随机抽取一个，设为 i_1, 令 $B=(i_1)$；然后把 i_1 从 A 中剔除，从剩下的 $n-1$ 个元素中继续随机抽取一个，设为 i_2，令 $B=(i_1,i_2)$，如此重复直至 A 中的元素都被抽取到 B 中。

这种做法需要两个向量 A 和 B 来存储。为了减少存储的使用，A 中仅剩 k 个元素时，向量 A 的后面 $n-k$ 个位置可以用来存放已经抽取出来的元素。从 A 的前 k 个中随机抽取一个后，可以把这个元素放在 A_k 的位置并把原来 A_k 填在抽取产生的空位。

$x \leftarrow (1,2,\ldots,n)$
$k \leftarrow n$
while $(k>1)$ {
  生成 $U\sim \mathrm{U}(0,1)$, 令 $I \leftarrow \mathrm{ceil}(kU)$
  交换 x_I 与 x_k
  $k \leftarrow k-1$
}
输出当前的向量 x 为随机排列结果

这个算法中 x 的初值也可以取为 $(1,2,\ldots,n)$ 的任何一个排列。另外，如果只需要从 $(1,2,\ldots,n)$ 中随机无放回地抽取 r 个 $(1\le r\le n)$，上面的算法可以很容易地改成

$x \leftarrow (1,2,\ldots,n)$
$k \leftarrow n$
while $(k>n-r+1)$ {
  生成 $U\sim \mathrm{U}(0,1)$, 令 $I \leftarrow \mathrm{ceil}(kU)$
  交换 x_I 与 x_k
  $k \leftarrow k-1$
}
输出 $(x_{n-r+1},x_{n-r+2},\ldots,x_n)$

随机无放回抽取 r 个可以假设 $r\le \frac{n}{2}$, 否则只要抽取余集就可以了。

生成随机排列的另一种做法是生成 n 个 U(0,1) 随机数 $(U_1,U_2,\ldots,U_n)$，求得把 $(U_1,U_2,\ldots,U_n)$ 从小到大排列的下标次序 $(i_1,i_2,\ldots,i_n)$，则 $(i_1,i_2,\ldots,i_n)$ 是 $(1,2,\ldots,n)$ 的一个随机排列。这种作法看起来比较巧妙，但是作排序需要执行 $n\log(n)$ 次比较运算，比上面的每次挑选一个的算法计算量大。

注：R 中用 `sample(x, size=n)` 从向量 `x` 中随机无放回地抽取 `n` 个。

例 2.2.4 (几何分布随机数). 设随机变量 X 表示在成功概率为 $p(0<p<1)$ 的独立重复试验中首次成功所需的试验次数，则 X 的概率分布为

$$P(X=k)=pq^{k-1}, k=1,2,\ldots,\ (q=1-p)$$

称 X 服从几何分布, 记为 $X\sim \mathrm{Geom}(p)$。

设 $U\sim \mathrm{U}(0,1)$，注意到

$$\begin{aligned} F(k) =& P(X\le k)=P(\text{试验在前 } k \text{ 次中有成功})\\ =& 1-P(\text{试验在前 } k \text{ 次中都失败})\\ =& 1-q^k,\ k=1,2,\ldots \end{aligned}$$

利用定理2.2.2, 生成 X 的方法为当且仅当 $1-q^{k-1}<U\le 1-q^k$ 时取 $X=k$，$k=1,2,\ldots$。此条件等价于

$$q^k\le 1-U<q^{k-1}$$

取

$$\begin{aligned} X =& \min\{k: q^k\le 1-U\}\\ =& \min\{k: k\log(q)\le \log(1-U)\}\\ =& \min\{k: k\ge \frac{\log(1-U)}{\log(q)}\}\\ =& \mathrm{ceil}\left(\frac{\log(1-U)}{\log(q)}\right) \end{aligned}$$

在没有指定底数时，$\log(\cdot)$ 默认使用自然对数。注意到 $1-U$ 也是服从 U(0,1) 分布的，所以只要取

$$X=\mathrm{ceil}\left(\frac{\ln(U)}{\ln(q)}\right).$$

则 X 服从几何分布。

例 2.2.5 (独立试验序列). 产生 $X_1,X_2,\ldots,X_N$ iid b(1,p)，即

$$P(X_i=1)=p=1-P(X_i=0),\ i=1,2,\ldots,N$$

设 $(U_1, U_2, \ldots, U_N)$ iid U(0,1), 则当 $U_i \le p$ 时取 $X_i = 1$, 当 $U_i > p$ 时取 $X_i = 0$ 可以构造独立试验序列 $(X_1, X_2, \ldots, X_n)$。

当 $0 < p < \frac{1}{2}$ 时利用例2.2.4可以更快地构造独立试验序列。想法是生成首次成功时间，则在成功之前的试验都是失败，然后再生成下次成功时间，两次成功之间的试验为失败，一直到凑够 N 次试验为止。算法如下：

$k \leftarrow 0$ (k 表示已经生成的个数)
while($k \le N$) {
  生成 $T \sim \text{Geom}(p)$
  if($T > 1$) {
    把 $X_{k+1}, X_{k+2}, \ldots, X_{\min(k+T-1,N)}$ 都赋值为 0
  }
  if($k + T \le N$) $X_{k+T} \leftarrow 1$
  $k \leftarrow k + T$
}

当 $0 < p < \frac{1}{2}$ 的情况，这时失败多于成功，用较少的几何分布随机数就能产生较长的序列。当 $p > \frac{1}{2}$ 时，应该生成首次失败的几何分布随机变量，然后算法中也改为若干个成功后面有一个失败。

例 2.2.6 (二项分布随机数). 设 X 为 n 次成功概率为 p 的独立重复试验的成功次数，则

$$p_k = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, \ldots, n\ (0 < p < 1)$$

称 X 服从二项分布，记为 $X \sim \text{B}(n, p)$。$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ 是从 n 个中取出 k 个的不同组合的个数。

为了生成二项分布随机数，可以生成 n 个两点分布随机数（长度为 n 的独立试验序列），其和为一个二项分布随机数，依次从独立试验序列中每 n 个求和就可以得到二项分布随机数序列。这样的算法效率较差。

设法用定理2.2.2构造二项分布随机数。二项分布取值概率有如下递推式:

$$p_{k+1} = \frac{n-k}{k+1} \frac{p}{1-p} p_k,\ k = 0, 1, \ldots, n-1$$

所以在利用定理2.2.2构造二项分布随机数时，可以递推计算

$$F(k+1) = F(k) + p_{k+1} = F(k) + \frac{n-k}{k+1} \frac{p}{1-p} p_k,\ k = 0, 1, \ldots, n-1 \tag{2.2}$$

算法如下

生成 $U \sim \mathrm{U}(0,1)$
$k \leftarrow 0$, $c \leftarrow \frac{p}{1-p}$, $a \leftarrow (1-p)^n$, $F \leftarrow a$
while$(U > F)$ {
  $a \leftarrow \frac{n-k}{k+1} ca$, $F \leftarrow F + a$
  $k \leftarrow k+1$
}
$X \leftarrow k$

当 $p > \frac{1}{2}$ 时，可以用上述算法生成 $Y \sim \mathrm{B}(n, 1-p)$，令 $X = n - Y$ 则 $X \sim \mathrm{B}(n,p)$，可以减少判断次数。

上面的算法先判断 X 是否应该取 0，如不是再判断 X 是否应该取 1，……，判断次数为 X 的值加 1，所以平均判断次数为 $EX+1 = np+1$。当 np 较大时，仿照例2.2.2的讨论，应该先判断取值概率较大的那些值。$P(X=k)$ 在 np 附近达到最大值。令 $K = \mathrm{floor}(np)$，应先判断要生成的随机变量 X 是小于等于 K 还是大于 K。先用(2.2)计算出 $F(K)$。当 $U \le F(K)$ 时，X 取值小于等于 K，这时可依次判断 X 是否应取 $K, K-1, \ldots, 0$，在这个过程中可以反向递推计算各个 $F(k)$ 的值。当 $U > F(K)$ 时，X 取值大于 K，这时可依次判断 X 是否应取 $K+1, K+2, \ldots, n$，在这个过程中可以递推计算各个 $F(k)$ 的值。

改进后需要的判断次数约为 $|X - np| + 1$，因为 n 较大时二项分布近似正态分布 $\mathrm{N}(np, np(1-p))$，所以平均判断次数约为

$$\begin{aligned} & 1 + E|X - np| \\ =& 1 + \sqrt{np(1-p)} E \left| \frac{X-np}{\sqrt{np(1-p)}} \right| \\ \approx& 1 + \sqrt{np(1-p)} E|Z| \quad (\text{其中 } Z \sim \mathrm{N}(0,1)) \\ =& 1 + 0.798\sqrt{p(1-p)} \cdot \sqrt{n} \end{aligned}$$

当 $n \to \infty$ 时判断次数比原来的 $O(n)$ 降低到了 $O(\sqrt{n})$。

例 2.2.7 (泊松分布随机数). 设离散型随机变量 X 分布为

$$p_k = P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}, \ k = 0, 1, 2, \ldots (\lambda > 0)$$

称 X 服从泊松分布，记为 $X \sim \mathrm{Poisson}(\lambda)$。易见

$$p_{k+1} = \frac{\lambda}{k+1} p_k, \ k = 0, 1, 2, \ldots$$

所以在利用定理2.2.2构造泊松随机数时，可以递推计算

$$F(k+1) = F(k) + p_{k+1} = F(k) + \frac{\lambda}{k+1} p_k,\ k = 0, 1, 2, \ldots \tag{2.3}$$

算法如下：

生成 $U \sim \mathrm{U}(0,1)$
$k \leftarrow 0$, $p \leftarrow e^{-\lambda}$, $F \leftarrow p$
while$(U > F)$ {
  $p \leftarrow \frac{\lambda}{k+1} p$, $F \leftarrow F + p$
  $k \leftarrow k + 1$
}
$X \leftarrow k$

这样的算法先判断是否令 X 取 0，不成立则再判断是否令 X 取 1，如此重复，判断的次数为 X 的值加 1，所以平均判断次数需要 $EX + 1 = \lambda + 1$ 次，当 λ 较小时这种方法效率不错；如果 λ 比较大，这时 $p(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ 在最接近于 λ 的两个整数之一上最大，按照例2.2.2的讨论，应该先判断这两个整数及其邻近的点。令 $K = \mathrm{floor}(\lambda)$, 用(2.3)先递推计算出 $F(K)$ 和 $F(K+1)$，然后判断 $F(K) < U \le F(K+1)$ 是否成立，如果成立就令 $X = K + 1$; 否则，如果 $U \le F(K)$ 则依次判断是否应取 X 为 $K, K-1, \ldots, 0$, 在这个过程中可以反向递推计算各 $F(k)$；如果 $U > F(K+1)$ 则依次判断是否应取 X 为 $K+2, K+3, \ldots$，这个过程中仍然用(2.3)递推计算各 $F(k)$。

这样的改进的算法需要的判断次数约为 $|X - \lambda| + 1$，所以平均判断次数为 $E|X - \lambda| + 1$。因为 λ 较大时泊松分布渐近正态分布 $\mathrm{N}(\lambda, \lambda)$, 所以平均判断次数约为

$$\begin{aligned} 1 + E|X - \lambda| =& 1 + \sqrt{\lambda} E\left|\frac{X - \lambda}{\sqrt{\lambda}}\right| \\ \approx& 1 + \sqrt{\lambda} E|Z| \quad (\text{其中 } Z \sim \mathrm{N}(0,1)) \\ =& 1 + 0.798\sqrt{\lambda} \end{aligned}$$

当 λ 很大时平均判断次数由 $O(\lambda)$ 降低到了 $O(\sqrt{\lambda})$。

2.2.3 用变换方法生成连续型分布的随机数

连续型分布的随机数都可以用逆变换法生成：$X = F^{-1}(U)$, $U \sim \mathrm{U}(0,1)$。对于反函数 F^{-1} 容易计算的情形，逆变换法是最方便的。

例 2.2.8. Beta(2,1) 分布的分布密度和分布函数分别为

$$p(x) = 2x, x \in [0,1] \qquad F(x) = x^2, x \in [0,1]$$

密度函数为三角形。若 $U \sim \mathrm{U}(0,1)$，则 $X = \sqrt{U}$ 为 Beta(2,1) 随机数。

Beta(1,2) 分布的分布密度和分布函数分别为

$$p(x) = 2(1-x), x \in [0,1] \qquad F(x) = 1-(1-x)^2, x \in [0,1]$$

密度函数为三角形。若 $U \sim \mathrm{U}(0,1)$，则 $X = 1 - \sqrt{1-U}$ 为 Beta(1,2) 随机数, 因为 U 与 $1-U$ 同分布所以 $X = 1 - \sqrt{U}$ 也是 Beta(1,2) 随机数。

对于 $0 < m < 1$，三角形分布 Tri(0,1,m) 的密度函数为

$$p(x) = \begin{cases} \frac{2}{m}x, & 0 < x \le m \\ \frac{2}{1-m}(1-x), & m < x < 1 \end{cases}$$

分布函数为

$$F(x) = \begin{cases} \frac{x^2}{m}, & 0 < x \le m \\ 1 - \frac{(1-x)^2}{1-m}, & m < x < 1 \end{cases}$$

反函数为

$$F^{-1}(u) = \begin{cases} \sqrt{mu}, & 0 < u \le m \\ 1 - \sqrt{(1-m)(1-u)}, & m < u < 1 \end{cases}$$

所以若 $U \sim \mathrm{U}(0,1)$，则 $X = F^{-1}(U)$ 服从三角形分布 Tri(0,1,m)。

例 2.2.9 (指数分布随机数). 服从指数分布 Exp(λ) ($\lambda > 0$) 的随机变量 X 的分布密度和分布函数分别为

$$p(x) = \lambda e^{-\lambda x},\ x > 0$$
$$F(x) = 1 - e^{-\lambda x},\ x > 0.$$

于是反函数为

$$F^{-1}(u) = -\lambda^{-1}\log(1-u)$$

所以 $U \sim \mathrm{U}(0,1)$ 时 $X = -\lambda^{-1}\log(1-U)$ 服从 Exp(λ)。因为 $1-U$ 与 U 同分布，所以取 $X = -\lambda^{-1}\log U$ 也服从 Exp(λ)。

当 $\lambda = 1$ 时 Exp(1) 叫做标准指数分布，$-\log U$ 服从标准指数分布，标准指数分布乘以 λ^{-1} 即为 Exp(λ) 分布。

例 2.2.10. 利用指数分布和泊松过程的关系也可以生成泊松随机数。$N(t)$ 为到时刻 t 为止到来的事件个数 $(t \geq 0)$，如果两个事件到来之间的间隔都服从独立的 Exp(λ) 分布则 $N(t)$ 为泊松过程，且 $N = N(1)$ 服从参数为 λ 的泊松分布。若 $U_1, U_2, \ldots$ 是独立的 U(0,1) 随机变量列，$X_1, X_2, \ldots$ 是独立的 Exp(λ) 随机变量列，则

$$N = \max\{n : \sum_{i=1}^{n} X_i \leq 1\}$$

X_i 可以用 $-\lambda^{-1} \ln U_i$ 生成，所以

$$\begin{aligned} N &= \max\{n : -\lambda^{-1} \sum_{i=1}^{n} \ln U_i \leq 1\} \\ &= \max\{n : U_1 U_2 \ldots U_n \geq e^{-\lambda}\} \end{aligned}$$

可以这样生成泊松随机数：相继生成均匀随机数，直至其连乘积小于 $e^{-\lambda}$, 取 n 为使用的均匀随机数个数减 1，即

$$N = \min\{n : U_1 U_2 \ldots U_n < e^{-\lambda}\} - 1.$$

关于一元随机变量和二元随机向量变换后的分布，概率论中有如下定理。

定理 2.2.3. 设随机变量 X 有密度 $p(x)$，$Y = g(X)$ 是 X 的函数，函数 $g(\cdot)$ 有反函数 $x = g^{-1}(y) = h(y)$, $h(y)$ 有一阶连续导数，则 Y 有密度

$$f(y) = p(h(y)) \cdot |h'(y)|. \tag{2.4}$$

定理 2.2.4. 设随机向量 (X, Y) 具有联合密度函数 $p(x, y)$，令

$$\begin{cases} u = g_1(x, y), \\ v = g_2(x, y), \end{cases}$$

设 $(u, v) = (g_1(x, y), g_2(x, y))$ 的反变换存在唯一，记为

$$\begin{cases} x = h_1(u, v), \\ y = h_2(u, v), \end{cases}$$

设 h_1, h_2 的一阶偏导数存在，反变换的 Jacobi 行列式

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \neq 0,$$

则随机向量 (U,V) 的联合密度为

$$f(u,v)=p(h_1(u,v),h_2(u,v))\cdot|J|. \tag{2.5}$$

有些分布之间有已知的关系，可以利用这样的关系产生随机数。如:

- 二项分布 B(n,p) 是 n 个伯努利分布 b(1,p) 之和，可以产生 n 个伯努利分布 b(1,p) 随机数求和得到二项分布随机数。
- 如果 U 服从标准均匀分布 U(0,1), 则 $X=a+(b-a)U$ 服从 U(a,b)。
- 如果 Z 服从标准正态分布，则可以用 $X=\mu+\sigma Z$ 服从 N(μ,σ^2) 分布。
- 如果 Z 服从标准指数分布 Exp(1), 则 $X=\lambda^{-1}Z$ 服从指数分布 Exp(λ)。
- 如果 Z 服从标准伽马分布 Gamma(α,1), 则 $X=\lambda^{-1}Z$ 服从伽马分布 Gamma(α, λ)。
- n 个独立的 Exp(λ) 指数分布随机变量之和服从伽马分布 Gamma(n,λ)。设 $U_1,U_2,\ldots,U_n$ 为独立 U(0,1) 随机变量，令

$$\begin{aligned}X=&\sum_{i=1}^{n}(-\lambda^{-1})\ln U_i\\ =&-\lambda^{-1}\ln(U_1U_2\ldots U_n)\end{aligned}$$

则 X 服从 Gamma(n,λ) 分布。

- 若 $Z_1,Z_2,\ldots,Z_n$ 独立同标准正态分布，则 $X=Z_1^2+Z_2^2+\cdots+Z_n^2\sim\chi^2(n)$ 分布。
- 如果 $X\sim$ Gamma$(\frac{n}{2},1)$, 则 $2X\sim\chi^2(n)$。
- 若 $Y\sim$ N$(0,1)$, $Z\sim\chi^2(n)$ 且 Y 与 Z 相互独立，则 $X=Y/\sqrt{Z}$ 服从 t(n) 分布。
- 若 $Y\sim\chi^2(n_1)$, $Z\sim\chi^2(n_2)$, Y 与 Z 相互独立，则 $X=\frac{Y/n_1}{Z/n_2}\sim$ F(n_1,n_2)。

下面的定理给出了一种生成标准正态随机数的算法。

定理 2.2.5. 设 U_1,U_2 独立且都服从 U(0,1)，

$$\begin{cases}X=&\sqrt{-2\ln U_1}\cos(2\pi U_2),\\ Y=&\sqrt{-2\ln U_1}\sin(2\pi U_2)\end{cases} \tag{2.6}$$

则 X,Y 独立且都服从标准正态分布。(2.6)叫做 Box-Muller 变换。

证明. 从 (U_1, U_2) 到 (X, Y) 的逆变换为

$$\begin{cases} U_1 = \exp\left\{-\frac{1}{2}(X^2+Y^2)\right\} \\ U_2 = \frac{1}{2\pi}\tan^{-1}\frac{Y}{X} \end{cases}$$

逆变换的 Jacobi 行列式为

$$J = \begin{vmatrix} \frac{\partial X}{\partial U_1} & \frac{\partial X}{\partial U_2} \\ \frac{\partial Y}{\partial U_1} & \frac{\partial y}{\partial U_2} \end{vmatrix} = -\frac{1}{2\pi}\exp\left\{-\frac{1}{2}(X^2+Y^2)\right\}$$

所以 (X, Y) 的联合密度为

$$f(x,y) = 1 \cdot \frac{1}{2\pi}\exp\left\{-\frac{1}{2}(x^2+y^2)\right\} = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}$$

即 X, Y 独立且分别服从标准正态分布。 □

例 2.2.11 (用极坐标变换生成正态随机数). 只要生成两个独立的 U(0,1) 随机变量 U_1 和 U_2，按照定理2.2.5用(2.6) 就可以生成两个独立的标准正态分布随机数。

这种方法的另外一个版本参见习题13。 □

2.2.4 舍选法

设随机变量 Z 的分布函数很难求反函数 $F^{-1}(u)$。如果 Z 取值于有限区间 $[a, b]$ 且 $p(x)$ 有上界 M:

$$p(x) \le M, \ \forall x \in [a, b]$$

则可以用如下算法生成 Z 的随机数：

until $(Y \le p(X))$ {
  生成 $U_1, U_2 \sim \mathrm{U}(0,1)$
  取 $X \leftarrow a + (b-a)U_1$, $Y \leftarrow MU_2$
}
输出 $Z \leftarrow X$

这种方法叫做舍选法 I。如图2.1所示，每次循环生成的 (X, Y) 实际上构成了矩形 $[a, b] \times [0, M]$ 上的二维均匀分布，循环退出条件为 $Y \le p(X)$, 循环退出时 (X, Y) 的值落入了 $p(x)$ 曲线下方，这时 (X, Y) 服从曲边梯形 $\{(x, y) : a \le x \le b, 0 \le y \le p(x)\}$ 上的二维均匀分布，所以循环退出时 $Z = X$ 在 x 附近取值的概率是与 $p(x)$ 成正比的。

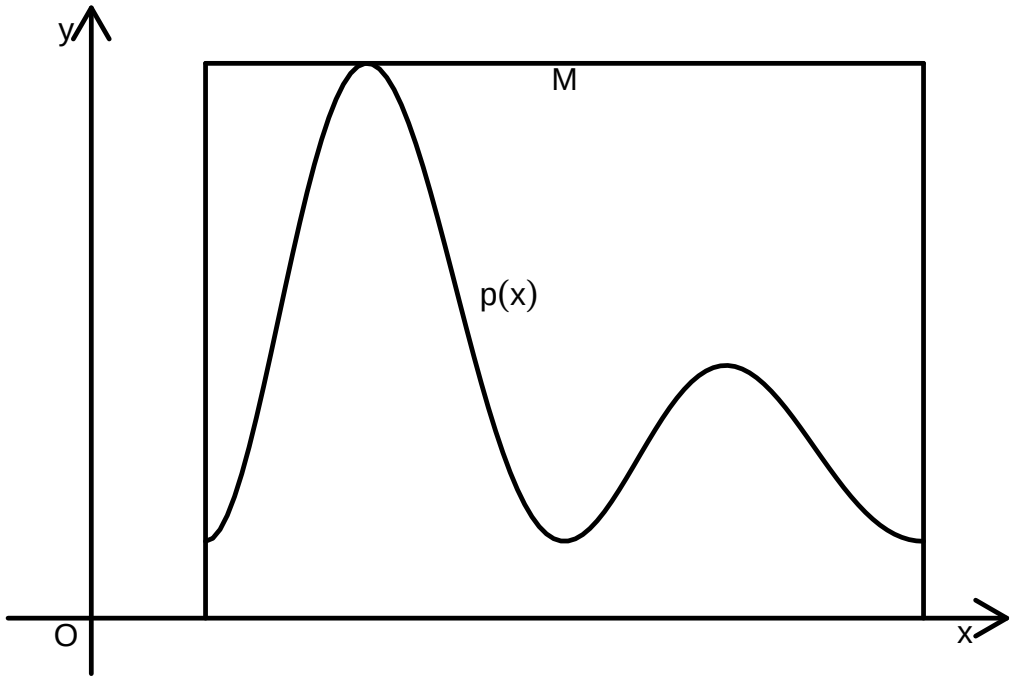

图 2.1: 舍选法 I 的示意图

定理 2.2.6. 舍选法 I 产生的 Z 密度为 $p(x)$，算法所需的迭代次数是均值为 $M(b-a)$ 的几何分布随机变量。

证明. 各次循环得到的 $(X_j, Y_j), j = 1, 2, \ldots$ 序列是独立同分布随机向量序列，循环结束条件为 $Y \le p(X)$, 所以循环次数 N 服从成功概率为 $p = P(Y \le p(X))$ 的几何分布，

$$\begin{aligned} p =& P(Y \le p(X)) = \int_a^b \int_0^{p(x)} \frac{1}{M(b-a)}\, dy\, dx \\ =& \frac{1}{M(b-a)}, \end{aligned}$$

算法结束时得到的 Z 的分布函数为

$$\begin{aligned}
F_Z(z) =& \sum_{j=1}^{\infty} P(X_j \le z \,|\, N = j) P(N = j) \\
=& \sum_{j=1}^{\infty} P(X_j \le z \,|\, Y_1 > p(X_1), \ldots, Y_{j-1} > p(X_{j-1}), Y_j \le p(X_j)) P(N = j) \\
=& \sum_{j=1}^{\infty} P(X_j \le z \,|\, Y_j \le p(X_j)) P(N = j) \quad \text{(由独立性)} \\
=& P(X_1 \le z \,|\, Y_1 \le p(X_1)) \sum_{j=1}^{\infty} P(N = j) \\
=& \int_a^z p(x)\, dx,
\end{aligned}$$

即 Z 的密度为 $p(x)$。 □

舍选法 I 的优点是只要随机变量在有限区间取值且密度函数有界就可以使用这种方法。但是，从图2.1和定理2.2.6可以看出这种舍选法的缺点：首先，X 的取值范围必须有界；其次，X 的密度 $p(x)$ 必须有界；第三，当算法中 X 取到 $p(x)$ 值很小的位置时，被拒绝的概率很大, X 被接受的概率实际是图2.1中曲线 $p(x)$ 下的面积和整个矩形面积之比，当 $p(x)$ 高低变化很大时这个比例很低，算法效率很差。

为了解决第一个和第二个问题，我们抽取 X 时不再从 $\mathrm{U}(a,b)$ 抽取，而可以从一个一般密度 $g(x)$ 抽取，称这个密度为“**试投密度**”，要求 $g(x)$ 的随机数容易生成。为了解决第三个问题，我们应该力图使得 $g(x)$ 形状与 $p(x)$ 相像，即 $p(x)$ 大时相应 $g(x)$ 也大，$p(x)$ 小时相应 $g(x)$ 也小，这样目标密度小的地方尝试抽取 X 也少。在很多实际问题中，目标密度 (要生成随机数的密度)$p(x)$ 可能本身是未知的，只知道 $\tilde{p}(x) = ap(x)$, 其中 a 为未知常数。要求存在常数 c 使得

$$\frac{\tilde{p}(x)}{g(x)} \le c, \ \forall x$$

这样，如果 $\tilde{p}(x)$ 在接近 $\pm\infty$ 处有定义，需要满足 $\tilde{p}(x) = O(g(x))(x \to \pm\infty)$。选取试投密度 $g(\cdot)$ 时上述常数 c 越小越好。

这时算法改为如下的舍选法 II:

until $(Y \le \frac{\tilde{p}(X)}{cg(X)})$ {
  生成 $X \sim g(x)$
  生成 $Y \sim U(0,1)$

}
输出 $Z \leftarrow X$

定理 2.2.7. 舍选法 II 产生的 Z 密度为 $p(x)$，算法所需的迭代次数是均值为 $\frac{c}{a}$ 的几何分布随机变量。

证明. 算法的循环中每次生成的 X 与 Y 独立，(X,Y) 的联合密度为

$$p(x,y) = g(x) \cdot 1 = g(x),\ \forall x \in (-\infty,\infty), y \in [0,1]$$

利用定理2.2.6的证明方法可知算法停止时 Z 的分布是 X 在 $Y \le \frac{\tilde{p}(X)}{cg(X)}$ 条件下的条件分布，所以输出的 Z 的分布函数为

$$\begin{aligned} F(x) =& P(Z \le x) = P\left(X \le x \,\middle|\, Y \le \frac{ap(X)}{cg(X)}\right) = \frac{P\left(X \le x,\ Y \le \frac{ap(X)}{cg(X)}\right)}{P\left(Y \le \frac{ap(X)}{cg(X)}\right)} \\ =& \frac{\int_{-\infty}^{x} g(u)du \int_0^{\frac{ap(u)}{cg(u)}} dv}{\int_{-\infty}^{\infty} g(u)du \int_0^{\frac{ap(u)}{cg(u)}} dv} = \int_{-\infty}^{x} p(u)du \end{aligned}$$

即输出的 Z 的分布密度为 $p(x)$。

算法所需的迭代次数 N 服从成功概率 $p = P\left(Y \le \frac{\tilde{p}(X)}{cg(X)}\right)$ 的几何分布，

$$p = P\left(Y \le \frac{ap(X)}{cg(X)}\right) = \int_{-\infty}^{\infty} \left(\int_0^{\frac{ap(x)}{cg(x)}} 1\, dy\right) g(x)\, dx = \frac{a}{c}$$

所以平均迭代次数为 $\frac{c}{a}$。 □

从定理可以看出，为了提高舍选法 II 的效率，应该取试投密度 $g(x)$ 使得 $g(x)$ 与 $p(x)$ 形状越接近越好, 常数 c 越小越好。

例 2.2.12. 用舍选法产生 Beta(2,4) 的随机数，密度为

$$p(x) = 20x(1-x)^3,\ 0 < x < 1.$$

可以用舍选法 I，这时

$$M = \max_{0 \le x \le 1} p(x) = p(\frac{1}{4}) = \frac{135}{64},$$

算法为

until $(Y \le 20X(1-X)^3)$ {
  生成 $U_1, U_2 \sim \mathrm{U}(0,1)$
  取 $X \leftarrow U_1$, $Y \leftarrow \frac{135}{64}U_2$
}
输出 $Z \leftarrow X$

平均迭代次数为 $\frac{135}{64} \approx 2.1$。

例 2.2.13. 生成 Gamma$(\frac{3}{2}, 1)$ 的随机数 Z，密度为

$$p(x) = \frac{1}{\Gamma(\frac{3}{2})} x^{\frac{1}{2}} e^{-x},\ x > 0$$

其中 $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$。用舍选法 II。$p(x)$ 形状在尾部与指数分布相近，且 $EZ = \frac{3}{2}$，所以使用期望为 $\frac{3}{2}$ 的指数分布作为试投密度 $g(x) = \frac{2}{3}e^{-\frac{2}{3}x}$, 求常数 c 使得 $\sup(p(x)/g(x)) \le c$。用微分法求得 $p(x)/g(x)$ 最大值点 $\frac{3}{2}$，最大值为 $c = \frac{3\sqrt{3}}{\sqrt{2\pi e}}$, 这时

$$\frac{p(x)}{cg(x)} = \sqrt{\frac{2e}{3}} x^{\frac{1}{2}} e^{-\frac{1}{3}x},$$

舍选法 II 的算法为:

until $(Y \le \sqrt{\frac{2e}{3}} X^{\frac{1}{2}} e^{-\frac{1}{3}X})$ {
  生成 $U \sim \mathrm{U}(0,1)$, $X \leftarrow -\frac{3}{2}\ln U$
  生成 $Y \sim U(0,1)$
}
输出 $Z \leftarrow X$

平均迭代次数为 $c = \frac{3\sqrt{3}}{\sqrt{2\pi e}} \approx 1.257$.

例 2.2.14. 把例2.2.13推广到一般伽马分布。设 Z 服从伽马分布 Gamma(α, λ)，密度为

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x},\ x > 0\ (\alpha > 0, \lambda > 0)$$

Z 的期望值为 α/λ。为生成 Z 的随机数，使用指数分布 Exp(t) 作为试投密度 $g(x)$，如何取 t 呢?

令

$$h(x) = \frac{p(x)}{g(x)} = \frac{\lambda^\alpha}{t\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda-t)x},$$

t 需要使得 $h(x)$ 有上界且上界最小。当 $0<\alpha<1$ 时 $\lim_{x\to 0} h(x)=+\infty$ 所以 $0<\alpha<1$ 时不能使用指数分布作为试投密度。当 $\alpha=1$ 时伽马分布就是指数分布。所以考虑 $\alpha>1$ 时 t 的取法。用微分法求得 $h(x)$ 的最大值点为 $x_0=\frac{\alpha-1}{\lambda-t}$, 于是 $h(x)$ 的最大值为

$$c(t)=h(x_0)=\frac{\lambda^\alpha}{\Gamma(\alpha)}(\alpha-1)^{\alpha-1}e^{-(\alpha-1)}\frac{1}{t(\lambda-t)^{\alpha-1}},$$

为了求最小的 $c(t)$ 只要求 $t(\lambda-t)^{\alpha-1}$ 的最大值点，用微分法易得 $t=\frac{\lambda}{\alpha}=\frac{1}{EX}$ 时 $c(t)$ 最小，所以试投密度的期望与 $p(x)$ 期望相同。

当伽马分布的参数 $\alpha\in(0,1)$ 时，可以找到如下分为 $0<x\le 1$ 和 $x>1$ 两段的控制函数:

$$M(x)=\begin{cases}\frac{1}{\Gamma(\alpha)}x^{\alpha-1}, & 0<x\le 1,\\ \frac{1}{\Gamma(\alpha)}e^{-x}, & x>1\end{cases}$$

使得 $p(x)\le M(x)$，然后把 $M(x)$ 归一化为一个密度就可以作为试投密度。

例 2.2.15 (用舍选法产生正态分布随机数). 如果 $X\sim \mathrm{N}(0,1)$ 则 $Y=\mu+\sigma X\sim \mathrm{N}(\mu,\sigma^2)$，所以只要产生 N(0,1) 随机数 Z。用舍选法 II，标准正态分布密度在尾部正比于 $e^{-\frac{1}{2}x^2}$，用标准指数分布密度来作为试投密度 (可以像例2.2.14那样证明在指数分布中参数为 1 的指数分布是最优的)。因为指数分布是非负的，所以先产生 $|Z|$ 的随机数，易见 $|Z|$ 的密度函数为

$$p(x)=\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}x^2},\ x>0$$

取 $g(x)=e^{-x}, x>0,$

$$h(x)=\frac{p(x)}{g(x)}=\sqrt{\frac{2}{\pi}}e^{x-\frac{1}{2}x^2},\ x>0$$

为使 $h(x)$ 最大只要使 $x-\frac{1}{2}x^2$ 最大，所以 $h(x)$ 最大值点为 $x_0=1$，取

$$c=h(1)=\sqrt{\frac{2e}{\pi}},\quad \frac{p(x)}{cg(x)}=e^{-\frac{1}{2}(x-1)^2},$$

产生 $|Z|$ 随机数后只需要以各自 $\frac{1}{2}$ 的概率加上正负号，所以产生标准正态分布随机数的舍选法算法为:

until $(Y\le e^{-\frac{1}{2}(X-1)^2})$ {
  生成 $U_1\sim \mathrm{U}(0,1)$, 令 $X\leftarrow -\ln U_1$
  生成 $Y\sim U(0,1)$

}
生成 $U_2 \sim \mathrm{U}(0,1)$
if $(U_2 < 0.5)$ {
  $Z \leftarrow X$
} **else** {
  $Z \leftarrow -X$
}
输出 Z

算法迭代次数为 $c = \sqrt{\frac{2e}{\pi}} \approx 1.32$。

例 2.2.16. 例2.2.11直接用 Box-Muller 变换生成两个标准正态分布随机数需要计算正弦和余弦函数，比较耗时。在 Box-Muller 公式中令 $R^2 = \sqrt{-2\ln U_1}$, $\theta = 2\pi U_2$，则 R^2 与 θ 独立，$R^2 \sim \mathrm{Exp}(\frac{1}{2})$, $\theta \sim \mathrm{U}(0, 2\pi)$, R 是 (X,Y) 的极径，θ 是 (X,Y) 的极角，是一个均匀随机角度。我们只要能生成极角服从 $\mathrm{U}(0,2\pi)$ 的随机点的直角坐标就可以获得 $\cos\theta$ 和 $\sin\theta$ 的值而不需要计算正弦和余弦。

设随机向量 (V_1, V_2) 服从单位圆 $C = \{(x,y) : x^2 + y^2 \le 1\}$ 上的均匀分布 $\mathrm{U}(C)$，则 (V_1, V_2) 的极角 θ 服从 $\mathrm{U}(0, 2\pi)$。(V_1, V_2) 可以用舍选法产生。得到这样的 (V_1, V_2) 后只要令

$$X = \sqrt{-2\ln U_1}\frac{V_1}{\sqrt{V_1^2 + V_2^2}},\ Y = \sqrt{-2\ln U_1}\frac{V_1}{\sqrt{V_2^2 + V_2^2}}$$

则 X, Y 服从独立的标准正态分布。

另外还可以证明，若 (V_1, V_2) 服从单位圆上的均匀分布 $\mathrm{U}(C)$, 则 $S = V_1^2 + V_2^2$ 服从 U(0,1) 且与 (V_1, V_2) 的极角 θ 独立。这样，U_1 可以用 S 代替。改进的算法如下:

until$(S \le 1)$ {
  生成 $U_1, U_2 \sim \mathrm{U}(0,1)$
  令 $V_1 \leftarrow 2U_1 - 1$, $V_2 \leftarrow 2U_2 - 1$, $S \leftarrow V_1^2 + V_2^2$
}
令 $X \leftarrow \sqrt{\frac{-2\ln S}{S}}V_1$, $Y \leftarrow \sqrt{\frac{-2\ln S}{S}}V_2$
输出 X 和 Y 作为独立的标准正态随机数

舍选法尤其适用于模拟仅取值于一个特殊区域的随机变量或随机向量, 例2.2.16单位圆内均匀分布 (V_1, V_2) 的产生就是用了舍选法。下面给出另一个例子。

例 2.2.17. 生成取值大于 5 的 Gamma(2,1) 的随机数 Z。密度为

$$p(x)=\frac{xe^{-x}}{\int_5^\infty ue^{-u}du}=\frac{xe^{-x}}{6e^{-5}},\ x>5$$

一个显然的算法为:

until $(X>5)$ {
  产生 $X\sim$ Gamma(2,1)
}
输出 $Z\leftarrow X$

这样的算法舍去小于 5 的试投值比较多，接受概率只有约 0.04，效率很低。

仍使用舍选法 II，试投密度使用大于 5 的指数分布。因为 Gamma(2,1) 期望为 $2/1=2$，使用大于 5 的 Exp(1/2)，密度为

$$g(x)=\frac{e^{-\frac{1}{2}x}}{\int_5^\infty e^{-\frac{1}{2}u}du}=\frac{1}{2}e^{\frac{5}{2}}e^{-\frac{1}{2}x},\ x>5$$

比值

$$h(x)=\frac{p(x)}{g(x)}=\frac{1}{3}e^{\frac{5}{2}}xe^{-\frac{1}{2}x},\ x>5$$

是单调减函数，所以 $h(x)\le c=h(5)=\frac{5}{3}$,

$$\frac{p(x)}{cg(x)}=\frac{1}{5}e^{\frac{5}{2}}xe^{-\frac{1}{2}x}.$$

如何生成 $g(x)$ 的随机数呢？注意指数分布的随机变量 ξ 有如下的无记忆性：$P(\xi>a+b|\xi>a)=P(\xi>b)$，所以在 $\xi>a$ 条件下的指数分布，可以看成是从 a 出发的一个指数分布，即在 $\xi>a$ 条件下的指数分布与 $\xi+a$ 同分布。所以 $g(x)$ 的随机数可以用指数分布随机数加 5 实现。算法如下：

until $(Y\le\frac{1}{5}e^{\frac{5}{2}}Xe^{-\frac{1}{2}X})$ {
  生成 $U\sim$ U$(0,1)$, $X\leftarrow 5-2\ln U$
  生成 $Y\sim U(0,1)$
}
输出 $Z\leftarrow X$

2.2.5 复合法

设离散型随机变量 I 的概率分布为

$$P(I=i)=\alpha_i,\ i=1,2,\ldots,m$$

若随机变量 $Z_1,Z_2,\ldots,Z_m$ 都是离散型随机变量，定义随机变量 X 的值为

$$X=\begin{cases} Z_1, & \text{当}I=1,\\ Z_2, & \text{当}I=2,\\ \cdots & \cdots\cdots\\ Z_m, & \text{当}I=m \end{cases} \tag{2.7}$$

则 X 是离散型随机变量，且

$$P(X=x)=\sum_{i=1}^{m}P(X=x|I=i)P(I=i)=\sum_{i=1}^{m}\alpha_i P(Z_i=x)$$

如果 X 的分布可以这样定义，则可以先生成 I 的样本，再根据 I 的值生成 Z_I 的样本，结果就是 X 的样本。

如果上面的 $Z_1,Z_2,\ldots,Z_m$ 都是连续型随机变量，密度函数分别为 $p_1(z),p_2(z),\ldots,p_m(z)$，则 X 的分布函数为

$$\begin{aligned}F(x)=&P(X\le x)=\sum_{i=1}^{m}P(X\le x|I=i)P(I=i)\\ =&\sum_{i=1}^{m}\alpha_i P(Z_i\le x)\end{aligned}$$

于是 X 有分布密度

$$p(x)=F'(x)=\sum_{i=1}^{m}\alpha_i p_i(x)$$

如果 X 的分布密度有这样的形式，可以先生成 I，然后根据 I 的值生成 Z_I 作为 X。

复合法需要比较多的随机数，在其它方法很难实现或效率较低时复合法有它的优势。

例 2.2.18. 设离散型随机变量 X 的分布为

$$P(X=k)=\begin{cases}0.05, & j=1,2,3,4,5\\ 0.15, & j=6,7,8,9,10\end{cases}$$

虽然可以当作普通的取有限个值的离散型随机变量来生成，但是 X 的分布明显可以看成如下的混合分布：随机变量 I 分布为

$$P(V=1)=0.25,\ P(V=2)=0.75$$

随机变量 Z_1 服从 $1,2,3,4,5$ 上的离散均匀分布，随机变量 Z_2 服从 $6,7,8,9,10$ 上的离散均匀分布，X 的分布为 $I=1$ 时取 Z_1, $I=2$ 时取 Z_2。所以，生成 X 的随机数的算法如下：

生成两个独立的均匀分布随机数 U_1 和 U_2
if$(U_1 < 0.75)$ {
  $X \leftarrow \text{ceil}(5 * U_2) + 5$
} **else** {
  $X \leftarrow \text{ceil}(5 * U_2)$
}

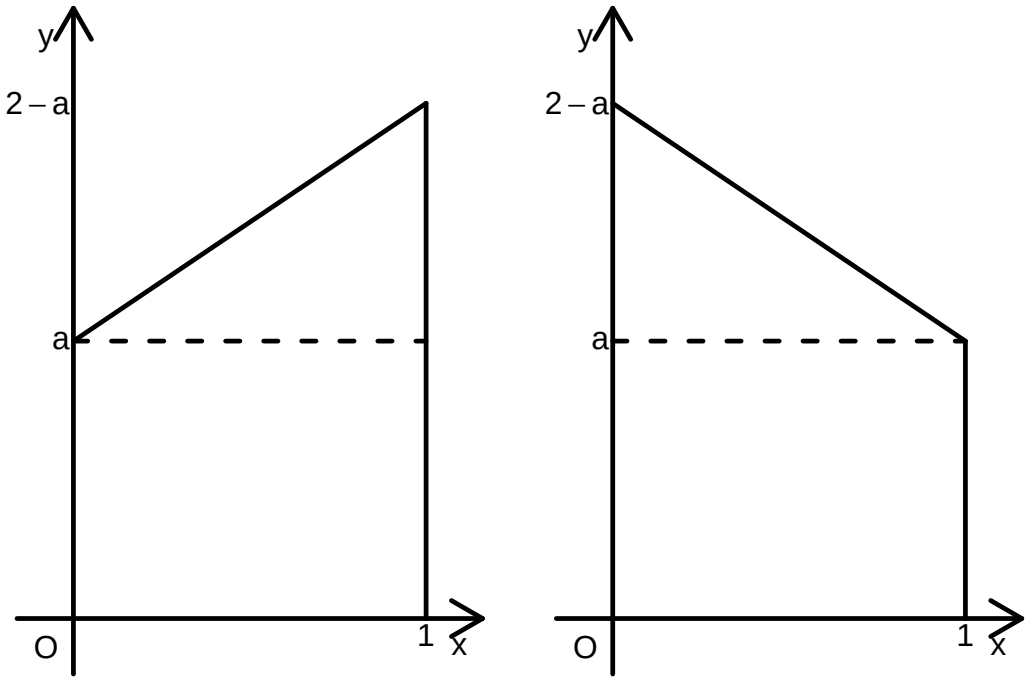

图 2.2: 梯形密度

例 2.2.19. 对 $0 < a < 1$，考虑梯形密度 (密度函数图像见图2.2)

$$p_1(x) = a + 2(1-a)x,\ 0 < x < 1,$$

或

$$p_2(x) = 2 - a - 2(1-a)x,\ 0 < x < 1$$

因为分布函数为

$$F_1(x) = ax + (1-a)x^2,\ 0 < x < 1$$

或

$$F_2(x) = (2-a)x - (1-a)x^2,\ 0 < x < 1$$

分布函数反函数为

$$F_1^{-1}(u) = \frac{-a + \sqrt{a^2 + 4(1-a)u}}{2(1-a)},\ 0 < u < 1$$

或

$$F_2^{-1}(u) = \frac{2 - a - \sqrt{(2-a)^2 - 4(1-a)u}}{2(1-a)},\ 0 < u < 1$$

可以用逆变换法获得 $p_1(x)$ 和 $p_2(x)$ 的随机数。另外，注意 $X \sim p_1(x)$ 当且仅当 $1-X \sim p_2(x)$，为了生成 $p_2(x)$ 的随机数，也可以生成 $X \sim p_1(x)$ 然后用 $1-X$ 作为 $p_2(x)$ 的随机数。

下面给出复合法生成 $p_1(x)$ 随机数的做法。$p_1(x)$ 下的区域可以分解为一个面积等于 a 的矩形和一个面积等于 $1-a$ 的三角形，令

$$\begin{aligned} g_1(x) &= 1,\ 0 < x < 1, \\ g_2(x) &= 2x,\ 0 < x < 1, \\ p_1(x) &= a \cdot g_1(x) + (1-a) \cdot g_2(x),\ 0 < x < 1 \end{aligned}$$

可见 $p_1(x)$ 是由一个均匀分布和一个三角形分布复合而成。于是，生成梯形 $p_1(x)$ 的随机数的复合法算法为

生成 $U_1 \sim \mathrm{U}(0,1)$

if$(U_1 < a)$ {
  生成 $X \sim \mathrm{U}(0,1)$
} **else** {
  生成 $U_2 \sim \mathrm{U}(0,1)$, 令 $X \leftarrow \sqrt{U}$
}
输出 X

在 R 软件中，提供了多种分布的随机数函数，同时也提供了这些分布的分布密度/分布概率函数、分布函数、分位数函数。比如，`rnorm(n, mu, sigma)` 生成 n 个期望为 `mu`、标准差为 `sigma` 的正态分布随机数, `dnorm(x, mu, sigma)` 返回期望为 `mu`、标准差为 `sigma` 的正态分布的密度函数在 $\boldsymbol{x}$ 处的函数值，`pnorm(x, mu, sigma)` 返回期望为 `mu`、标准差为 `sigma` 的正态分布的分布函数在 $\boldsymbol{x}$ 处的函数值, `qnorm(p, mu, sigma)` 返回期望为 `mu`、标准差为 `sigma` 的正态分布的分位数函数在 $\boldsymbol{p}$ 处的函数值。为了查看这些与分布有关的函数，在 R 命令行键入

```
?Distributions
```

将打开操作系统默认的互联网浏览器，显示 R 软件中与分布有关的函数的帮助信息。

2.3 随机向量和随机过程的生成

2.3.1 条件分布法

产生随机向量的一种方法是条件分布法。设 $\boldsymbol{X} = (X_1, X_2, \ldots, X_r)$ 的分布密度或分布概率 $p(x_1, x_2, \ldots, x_r)$ 可以分解为

$$p(x_1, x_2, \ldots, x_r) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_r|x_1, x_2, \ldots, x_{r-1})$$

则可以先生成 X_1，由已知的 X_1 的值从条件分布 $p(x_2|x_1)$ 产生 X_2，再从已知的 X_1, X_2 的值从条件分布 $p(x_3|x_1, x_2)$ 产生 X_3，如此重复直到产生 X_r。

例 2.3.1 (多项分布随机数). 进行了 n 次独立重复试验 $Y_1, Y_2, \ldots, Y_n$，每次的试验结果 Y_i 在 $1, 2, \ldots, r$ 中取值，$P(Y_i = j) = p_j, j = 1, 2, \ldots, r; i = 1, 2, \ldots, n$。令 X_j 为这 n 次试验中结果 j 的个数 $(j = 1, 2, \ldots, r)$，称 $\boldsymbol{X} = (X_1, X_2, \ldots, X_r)$ 服从多项分布，其联合概率函数为

$$P(X_j = x_j, j = 1, 2, \ldots, r) = \frac{n!}{x_1!x_2!\ldots x_r!}p_1^{x_1}p_2^{x_2}\ldots p_r^{x_r}.$$

(其中 $\sum_{j=1}^{r} x_j = n$)。

要生成 $\boldsymbol{X}$ 的随机数，考虑不同结果数 r 与试验次数 n 的比较。当 r 和 n 相比很大时，每个结果的出现次数都不多，而且许多结果可能根本不出现，这样，可以模拟产生 $Y_i, i = 1, 2, \ldots, n$ 然后从 $\{Y_i\}$ 中计数得到 $(X_1, X_2, \ldots, X_r)$。当 r 与 n 相比较小时，每个结果出现次数都比较多，可以使用条件分布逐个地产生 $X_1, X_2, \ldots, X_r$。

X_1 表示 n 次重复试验中结果 1 的出现次数，以结果 1 作为成功，其它结果作为失败，显然 $X_1 \sim \mathrm{B}(n, p_1)$。产生 X_1 的值 x_1 后，n 次试验剩余的 $n - x_1$ 次试验结果只有 $2, 3, \ldots, r$ 可取，于是在 $X_1 = x_1$ 条件下结果 2 的出现概率是

$$P(Y_i = 2 | Y_i \neq 1) = \frac{P(Y_i = 2)}{\sum_{k=2}^{r} P(Y_i = k)} = \frac{p_2}{1 - p_1}$$

于是剩下的 $n - x_1$ 次试验中结果 2 的出现次数服从 $\mathrm{B}(n - x_1, \frac{p_2}{1-p_1})$ 分布，从这个条件分布产生 $X_2 = x_2$。类似地，剩下的 $n - x_1 - x_2$ 次试验中结果 3 的出现次数服从 $\mathbf{B}(n - x_1 - x_2, \frac{p_3}{1-p_1-p_2})$ 分布，从这个条件分布中抽取 $X_3 = x_3$，如此重复可以产生多项分布 $\boldsymbol{X}$ 的随机数 $(x_1, x_2, \ldots, x_r)$。

2.3.2 多元正态分布模拟

设随机向量 $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ 服从多元正态分布 $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$，联合密度函数为

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}, \ \boldsymbol{x} \in R^p$$

正定矩阵 Σ 有 Cholesky 分解 $\Sigma = CC^T$，其中 C 为下三角矩阵 (见 §5.2.3)。设 $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_p)^T$ 服从 p 元标准正态分布 $\mathrm{N}(\mathbf{0}, I)$($I$ 表示单位阵)，则 $\boldsymbol{X} = \boldsymbol{\mu} + C\boldsymbol{Z}$ 服从 $\mathrm{N}(\boldsymbol{\mu}, \Sigma)$ 分布。

2.3.3 用 copula 描述多元分布 *

实际问题建模中，经常遇到两个随机变量的边缘分布很明确，但是它们之间的关系比较模糊，不能精确地用一个联合分布表示出来。这时，可以借助于 copula 分布来粗略地表示联合分布。

定义 (copula) 设随机向量 (X, Y) 有联合分布函数 $C(x, y)$, X 和 Y 的边缘分布都是标准均匀分布 U(0,1), 则称 (X, Y) 服从 copula 分布。copula 分布的概念也可以推广到 n 元随机向量情形。

如果 X 和 Y 都服从连续型分布，设 $X \sim F(x)$, $Y \sim G(y)$, 则 $F(X)$ 和 $G(Y)$ 都服从标准均匀分布，于是 $(F(X), G(Y))$ 服从 copula 分布。在不知道 (X, Y) 的真实联合分布情况

下，可以根据已知的 X,Y 的相关情况适当选取一个 copula 分布作为 $(F(X),G(Y))$ 的联合分布，反推得到 (X,Y) 的联合分布。

高斯 copula 是一个常用的 copula 分布。设 (X,Y) 服从二元联合正态分布，且其边缘分布为标准正态分布，相关系数为 ρ，则 $(\Phi(X),\Phi(Y))$ 服从 copula 分布，其分量的相关性可以用原来的 ρ 来刻画。高斯 copula 的联合分布函数为

$$\begin{aligned} C(x,y) =& P(\Phi(X)\le x,\Phi(Y)\le y) = P(X\le \Phi^{-1}(x), Y\le \Phi^{-1}(y)) \\ =& \int_{-\infty}^{\Phi^{-1}(x)}\int_{-\infty}^{\Phi^{-1}(y)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2+y^2-2\rho xy)\right\} dy\,dx. \end{aligned}$$

例 2.3.2. 设 $(X_1,X_2,\ldots,X_n)$ 服从多元正态分布，协方差阵为 Σ, 且各分量服从标准正态分布, 称 $(\Phi(X_1),\Phi(X_2),\ldots,\Phi(X_n))$ 服从多元高斯 copula 分布，设其分布函数为 $C(x_1,x_2,\ldots,x_n)$。

设随机向量 $(Z_1,Z_2,\ldots,Z_n)$ 的各分量分别有分布函数 $F_i(x), i=1,2,\ldots,n$，$F_i(x)$ 可逆。为了用多元高斯 copula 分布来模拟随机向量 $(Z_1,Z_2,\ldots,Z_n)$, 首先用 Cholesky 分解方法生成 $(X_1,X_2,\ldots,X_n)$ 的样本，令 $Y_i=\Phi(X_i), i=1,2,\ldots,n$, 则 $(Y_1,Y_2,\ldots,Y_n)$ 服从多元高斯 copula 分布。再令 $Z_i=F_i^{-1}(Y_i), i=1,2,\ldots,n$, 就得到了所需的 $(Z_1,Z_2,\ldots,Z_n)$ 的随机抽样。

2.3.4 泊松过程模拟 *

参数为 λ 的泊松过程 $N(t), t\ge 0$ 是取整数值的随机过程，$N(t)$ 表示到时刻 t 为止到来的事件个数，两次事件到来的时间间隔服从指数分布 $\mathrm{Exp}(\lambda)$。

所以，如果要生成泊松过程前 n 个事件到来的时间，只要生成 n 个独立的 $\mathrm{Exp}(\lambda)$ 随机数 $X_1,X_2,\ldots,X_n$，则 $S_k=\sum_{i=1}^k X_i, k=1,2,\ldots,n$ 为各个事件到来的时间。

如果要生成泊松过程在时刻 T 之前的状态，只要知道发生在 T 之前的所有事件到来时间就可以了。算法如下：

$S\leftarrow 0$, $A=\{\}$
repeat {
  生成 $U\sim \mathrm{U}(0,1)$, $X\leftarrow -\lambda^{-1}\ln U$
  $S\leftarrow S+X$
  if$(S\le T)$ {
    $A\leftarrow A\cup\{S\}$
  } **else** {
    break

}
}
输出集合 A

生成泊松过程在时刻 T 之前的状态的另外一种方法是先生成 $N(T) \sim \text{Poisson}(\lambda)$, 设 $N(T)$ 的值为 n，再生成 n 个独立的 U(0,1) 随机变量 $U_1, U_2, \ldots, U_n$，从小到大排序为 $U_{(1)} \le U_{(2)} \le \ldots U_{(n)}$, 则 $(TU_{(1)}, TU_{(2)}, \ldots, TU_{(n)})$ 为时刻 T 之前的所有事件到来时间（见习题14）。

为了生成强度函数为 $\lambda(t), t \ge 0$ 的非齐次泊松过程到时刻 T 为止的状态，如果 $\lambda(t)$ 满足

$$\lambda(t) \le M, \ \forall t \ge 0$$

则可以按照生成参数为 M 的齐次泊松过程的方法去生成各个事件到来时刻，但是以 $\lambda(t)/M$ 的概率实际记录各个时刻。算法如下:

$S \leftarrow 0$, $A = \{\}$
repeat {
  生成 $U_1 \sim \text{U}(0,1)$, $X \leftarrow -M^{-1} \ln U_1$
  $S \leftarrow S + X$
  if$(S \le T)$ {
    生成 $U_2 \sim \text{U}(0,1)$
    if$(U_2 \le \lambda(S)/M)$ $A \leftarrow A \cup \{S\}$
  } **else** {
    break
  }
}
输出集合 A

这种方法叫做稀松法 (thinning), 当 $\lambda(t)$ 变化不大时效率较高。如果 $\lambda(t)$ 变化很大，可以分段模拟，详见 Ross(2013)[34] §5.5。

2.3.5 平稳时间序列模拟 *

平稳时间序列中的 ARMA 模型可以递推生成。

对 AR(p) 模型

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \varepsilon_t, \ t \in \mathbb{Z} \tag{2.8}$$

($\mathbb{Z}$ 表示所有整数的集合), 其中 $\{\varepsilon_t\}$ 为白噪声 WN(0,σ^2), 即方差都为 σ^2、彼此不相关的随机变量序列，$\{\varepsilon_t\}$ 可以用 N(0,σ^2) 分布的独立序列来模拟，也可以用其它有二阶矩的分布。因为 AR(p) 模型具有所谓“稳定性”，所以我们从任意初值出发按照(2.8)递推 N_0 步 (N_0 是一个较大正整数) 后，再继续递推 N 步后得到的 N 个 X_t 就可以作为上述 AR(p) 模型的一次实现。根据模型稳定性的好坏，N_0 可取为 $50 \sim 1000$ 之间。算法为:

生成 $\{\varepsilon_t, t = 1, 2, \ldots, N_0 + N\}$ 服从独立 N(0,σ^2) 分布
置 $X_{-p+1}, X_{-p+2}, \ldots, X_0$ 都等于 0
for(t **in** 1:($N_0 + N$)) {
  $X_t \leftarrow a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \varepsilon_t$
}
输出 $(X_{N_0+1}, X_{N_0+2}, \ldots, X_{N_0+N})$

对于 MA(q) 模型

$$X_t = \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q},\ t \in \mathbb{Z} \tag{2.9}$$

生成 $\varepsilon_{1-q}, \varepsilon_{2-q}, \ldots, \varepsilon_0, \varepsilon_1, \ldots, \varepsilon_N$ 后可以直接对 $t = 1, 2, \ldots, N$ 按公式(2.9) 计算得到 $\{X_t, t = 1, 2, \ldots, N\}$。

对于 ARMA(p, q) 模型

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p} + \varepsilon_t + b_1\varepsilon_{t-1} + \cdots + b_q\varepsilon_{t-q},\ t \in \mathbb{Z} \tag{2.10}$$

也可以从初值零出发递推生成 $N_0 + N$ 个然后只取最后的 N 个作为 ARMA(p, q) 模型的一次实现。

R 函数 `filter` 可以进行递推计算。R 函数 `arima.sim` 可以进行 ARMA 模型和 ARIMA 模型模拟。

习题二

1. 对线性同余发生器

$$x_n = (ax_{n-1} + c) \pmod{M},\ n = 1, 2, \ldots$$

证明若从初值 x_0 出发周期等于 M，则以 $0 \sim M-1$ 中任何一个整数作为初值时周期都等于 M；若从某初值 x_0 出发周期小于 M，则以 $0 \sim M-1$ 中任何一个整数作为初值时周期都小于 M。

2. 写程序生成 n 个如下离散型分布随机数: $P(X=1)=1/3$, $P(X=2)=2/3$。

   (1) 生成 $n=100$ 个这样的随机数，计算 X 取 1 的百分比；

   (2) 生成 $n=1000$ 个这样的随机数，计算 X 取 1 的百分比；

   (3) 生成 $n=10000$ 个这样的随机数，计算 X 取 1 的百分比；

   (4) 用中心极限定理推导 n 个这样的随机数取 1 的百分比 f_n 的渐近分布，并用此渐近分布检验上面的三组样本是否与 $P(X=1)=1/3$ 相符。

3. 设随机变量 X 服从离散分布 $P(X=k)=p_k, k=1,2,\ldots,m$。编写 R 程序，输入 $\{p_k\}$，输出 n 个该离散分布的随机数。

4. 洗好一副编号分别为 1,2,...,54 的纸牌，依次抽取出来，若第 i 次抽取到编号 i 的纸牌则称为成功抽取。编写程序估计成功抽取个数 T 的期望和方差，推导理论公式并与模拟结果进行比较。

5. 做投掷两枚骰子的试验，连续试验直到点数之和的所有可能值 2,3,...,12 都出现一次，所需的试验次数记为 T，用随机模拟方法估计 T 的期望和方差。

6. 编写 R 程序，对某个均匀分布随机数发生器产生的序列做均匀性的卡方检验。

7. 设随机变量 X 密度函数 $p(x)$ 和分布函数 $F(x)$ 已知，编写 R 程序，对 X 的某个随机数发生器产生的序列做拟合优度卡方检验。

8. 编写例2.2.6中原始的和改进的生成二项分布随机数的算法并用 R 程序实现，比较两种算法得到的序列是否相同。

9. 编写例2.2.7中原始的和改进的生成泊松分布随机数的算法并用 R 程序实现，比较两种算法得到的序列是否相同。

10. 设随机变量 X 表示成功概率为 p 的独立重复试验中第 r 次成功所需要的试验次数，称 X 服从负二项分布。

    (1) 利用负二项分布与几何分布的关系构造 X 的随机数。

    (2) 直接利用负二项分布的概率分布构造 X 的随机数。

11. 用变换法生成如下分布的随机数:

(1) Beta($\frac{1}{n}$, 1) 分布，密度为

$$p(x) = \frac{1}{n} x^{\frac{1}{n}-1}, x \in [0,1]$$

(2) Beta(n, 1) 分布，密度为

$$p(x) = nx^{n-1}, x \in [0,1]$$

(3) 密度为

$$p(x) = \frac{2}{\pi\sqrt{1-x^2}}, x \in [0,1]$$

(4) 柯西分布，密度为

$$p(x) = \frac{1}{\pi(1+x^2)}, x \in (-\infty, \infty)$$

(5) 密度为

$$p(x) = \cos(x), x \in [0, \frac{\pi}{2}]$$

(6) 威布尔分布，密度为

$$p(x) = \frac{\alpha}{\eta} x^{\alpha-1} e^{-\frac{x^\alpha}{\eta}},\ x > 0\ (\alpha > 0, \eta > 0)$$

12. 设 X 密度为

$$p_R(x) = \frac{2(x-a)}{(b-a)^2},\ x \in (a,b)$$

此密度单调上升，是以 $[a,b]$ 为底的三角形，叫做 RT(a,b) 分布；若 X 的密度为

$$p_L(x) = \frac{2(b-x)}{(b-a)^2},\ x \in (a,b)$$

此密度单调下降，也是以 $[a,b]$ 为底的三角形，叫做 LT(a,b) 分布。试证明：

(1) 若 $X \sim \mathrm{RT}(0,1)$，则 $Y = a + (b-a)X \sim \mathrm{RT}(a,b)$; 若 $X \sim \mathrm{LT}(0,1)$，则 $Y = a + (b-a)X \sim \mathrm{LT}(a,b)$。

(2) $X \sim \mathrm{RT}(0,1)$ 当且仅当 $1 - X \sim \mathrm{LT}(0,1)$。

(3) 若 U_1, U_2 独立且分别服从 U(0,1) 分布，则 $X = \max(U_1, U_2) \sim \mathrm{RT}(0,1)$, $Y = \min(U_1, U_2) \sim \mathrm{LT}(0,1)$。

13. 设 $\alpha \sim \mathrm{U}(0, 2\pi)$, $R \sim \mathrm{Exp}(\frac{1}{2})$ 与 α 独立，令

$$\begin{cases} X = \sqrt{R}\cos\alpha \\ Y = \sqrt{R}\sin\alpha \end{cases}$$

证明 X, Y 相互独立且都服从 N(0,1) 分布。

14. 设 $U_1, U_2, \ldots, U_k, V_1, V_2, \ldots, V_{k-1}$ 独立同 U(0,1) 分布。令 $T = -\log(U_1 U_2 \ldots U_k)$, 设 $V_1, V_2, \ldots, V_{k-1}$ 从小到大排列结果为 $V_{(1)} \le V_{(2)} \le \ldots V_{(k-1)}$, 记 $V_{(0)} = 0$, $V_{(k)} = 1$, 令

$$X_i = T(V_{(i)} - V_{(i-1)}),\ i = 1, 2, \ldots, k,$$

证明 $X_1, X_2, \ldots, X_k$ 独立同标准指数分布 Exp(1)。

15. 设 X 为标准指数分布 Exp(1) 随机变量，模拟在 $X < 0.05$ 条件下 X 的分布，密度为

$$p(x) = \frac{e^{-x}}{1 - e^{-0.05}},\ 0 < x < 0.05$$

生成 1000 个这样的随机数，并用它们估计 $E(X|X < 0.05)$，推导 $E(X|X < 0.05)$ 的精确值并与模拟结果比较。

16. 证明定理2.2.5。

17. 设随机向量 (X, Y) 服从单位圆 $C = \{(x, y) : x^2 + y^2 \le 1\}$ 上的均匀分布，$R^2 = X^2 + Y^2$，θ 为 (X, Y) 的极角，则 R^2 与 θ 独立，$R^2 \sim \mathrm{U}(0,1)$, $\theta \sim \mathrm{U}(0, 2\pi)$。

18. 设随机变量 X 分布为

$$P(X = k) = \frac{e^{-\lambda}\frac{\lambda^k}{k!}}{\sum_{i=0}^{m} e^{-\lambda}\frac{\lambda^i}{i!}},\ k = 0, 1, \ldots, m$$

给出模拟此分布的两种方法。

19. 设 $X \sim \mathrm{B}(n, p)$, k 为满足 $0 \le k \le n$ 的给定的整数，随机变量 Y 的分布函数为 $P(Y \le y) = P(X \le y | X \ge k)$。记 $\alpha = P(X \ge k)$。分别用逆变换法和舍选法生成 Y 的随机数。当 α 取值大的时候还是取值小的时候舍选法不可取？

20. 设随机变量 X 分布函数为 $G(x)$，密度函数为 $g(x)$。对 $a < b$，令

$$F(x) = \frac{G(x) - G(a)}{G(b) - G(a)}, a \le x \le b$$

(1) $F(x)$ 是一个分布函数，其对应的分布是 X 在什么条件下的条件分布？

(2) 证明可以用如下方法生成 $F(x)$ 的随机数: 反复生成 $X \sim G(x)$，直到 $X \in [a,b]$，输出 X 的值为 $F(x)$ 的随机数。

21. 给出生成如下密度

$$p(x) = xe^{-x},\ x > 0$$

的随机数的两种方法并比较其效率。

22. 设随机变量分别以概率 0.06, 0.06, 0.06, 0.06, 0.06, 0.15, 0.13, 0.14, 0.15, 0.13 取值 1,2,...,10，应用复合方法给出产生此分布随机数的算法。

23. 设随机变量 X 的分布为

$$P(X = k) = \frac{1}{2^{k+1}} + \frac{2^{k-1}}{3^k},\ k = 1, 2, \ldots$$

给出模拟此随机变量的算法。

24. 设随机变量 X 的分布密度为

$$p(x) = \frac{1}{2}e^{-|x|},\ -\infty < x < \infty$$

称 X 服从双指数分布或 Laplace 分布。分别用逆变换法和复合法生成 X 的随机数。

25. 设随机变量 X 的分布密度为

$$p(x) = \frac{1}{2} + x,\ 0 \le x \le 1,$$

分别给出用逆变换法、舍选法、复合法生成 X 随机数的算法，并比较三种方法的效率。

26. 给出生成密度函数为

$$p(x) = \frac{1}{0.000336}x(1-x)^3, 0.8 < x < 1$$

的随机数的有效算法。

27. 设坛子中有 n 个不同颜色的球，第 i 中颜色的球有 n_i 个 $(i = 1, 2, \ldots, r)$。从坛子中随机无放回地抽取 m 个球，设随机变量 X_i 表示取出的第 i 种颜色球的个数，设计高效的算法模拟 $(X_1, X_2, \ldots, X_r)$ 的值。

28. 设随机变量 X 的概率分布为 $p_k = P(X = k), k = 1, 2, \ldots, \sum_{k=1}^{\infty} p_k = 1$。记

$$\lambda_n = P(X = n | X > n - 1) = \frac{p_n}{1 - \sum_{k=1}^{n-1} p_k},\ n = 1, 2, \ldots$$

(1) 证明 $p_1 = \lambda_1$, $p_n = (1 - \lambda_1)(1 - \lambda_2) \cdots (1 - \lambda_{n-1})\lambda_n$ $(n = 2, 3, \ldots)$。如果把 X 看成某种产品的寿命，则 λ_n 表示此产品在其寿命大于 $n - 1$ 的条件下，寿命为 n 的概率。$\{\lambda_n, n \geq 1\}$ 叫做**离散机会比**。一个生成 X 的随机数的方法是连续生成均匀随机数直到第 n 个小于 λ_n，叫做离散机会比方法，算法如下:

$k \leftarrow 0$
until $(U < \lambda_k)$ {
  生成 $U \sim \mathrm{U}(0, 1)$
  $k \leftarrow k + 1$
}
$X \leftarrow k$

(2) 证明上述算法产生的随机数符合 X 的分布。

(3) 假设 X 是一个参数为 p 的几何分布随机变量，求 $\lambda_n, n = 1, 2, \ldots$。说明此时上述算法是用来做什么的，它的有效性为什么是明显的?

29. 假设 X 为取值于 $\{1, 2, \ldots\}$ 的随机变量，$\lambda_n, n = 1, 2, \ldots$ 为其离散机会比，满足 $0 \leq \lambda_n \leq \lambda, n = 1, 2, \ldots$。用如下算法产生 X 的随机数:

$S \leftarrow 0$
until$(U_2 \leq \lambda_S/\lambda)$ {
  产生 $U_1 \sim \mathrm{U}(0, 1)$ 且令 $Y \leftarrow \mathrm{ceil}\left(\frac{\log(U_1)}{\log(1-\lambda)}\right)$
  $S \leftarrow S + Y$
  产生 $U_2 \sim \mathrm{U}(0, 1)$
}
$X \leftarrow S$

(1) 算法中的 Y 服从什么分布?

(2) 证明这样得到的 X 是离散机会比为 $\{\lambda_n\}$ 的离散型随机变量。

30. 利用稀松法编写模拟到时刻 $T = 10$ 为止的强度为

$$\lambda(t) = 3 + \frac{4}{t + 1}$$

的非齐次泊松过程的 R 程序; 设法改进这个算法的效率。

31. 用 R 的 `filter` 函数编写 AR(p) 模型的模拟程序。

32. 用 R 的 `filter` 函数编写 MA(q) 模型的模拟程序。

第三章　随机模拟

3.1　概述

在用数学模型，包括概率统计模型处理实际应用中的问题时，我们希望建立的模型能够尽可能地符合实际情况。但是，实际情况是错综复杂的，如果一味地要求模型与实际完全相符，会导致模型过于复杂，以至于不能进行严格理论分析，结果导致模型不能使用。所以，实际建模时会忽略许多细节，增加一些可能很难验证的理论假设，使得模型比较简单，可以用数学理论进行分析研究。

这样，简化的模型就可以与实际情况有较大的差距，即使我们对模型进行了完美的理论分析，也不能保证分析结果是可信的。这一困难可以用随机模拟的方法解决。

模拟是指把某一现实的或抽象的系统的某种特征或部分状态，用另一系统（称为模拟模型）来代替或模拟。为了解决某问题，把它变成一个概率模型的求解问题，然后产生符合模型的大量随机数，对产生的随机数进行分析从而求解问题，这种方法叫做**随机模拟**方法，又称为蒙特卡洛 (Monte Carlo) 方法。

例如，一个交通路口需要找到一种最优的控制红绿灯信号的办法，使得通过路口的汽车耽搁的平均时间最短，而行人等候过路的时间不超过某一给定的心理极限值。十字路口的信号共有四个方向，每个方向又分直行、左转、右转。因为汽车和行人的到来是随机的，我们要用随机过程来描述四个方向的汽车到来和路口的行人到来过程。理论建模分析很难解决这个最优化问题。但是，我们可以采集汽车和行人到来的频率，用随机模拟方法模拟汽车和行人到来的过程，并模拟各种控制方案，记录不同方案造成的等待时间，通过统计比较找出最优的控制方案。

随机模拟中的随机性可能来自模型本身的随机变量，比如上面描述的汽车和行人到来，也可能是把非随机的问题转换为概率模型的特征量估计问题从而用随机模拟方法解决。

例 3.1.1. 为了计算圆周率 π 的近似值可以使用随机模拟方法。如果向正方形 $D = \{(x, y) : x \in [-1, 1], y \in [-1, 1]$ 内随机等可能投点，落入单位圆 $C = \{(x, y) : x^2 + y^2 \leq 1\}$ 的概率为面

积之比 $p=\frac{\pi}{4}$。如果独立重复地投了 N 个点，落入 C 中的点的个数 ξ 的平均值为 $E\xi=pN$，由概率的频率解释，

$$\frac{\xi}{N}\approx\frac{\pi}{4},\quad \pi\approx\hat{\pi}=\frac{4\xi}{N}$$

可以这样给出 π 的近似值。 □

随机模拟方法会引入所谓**随机模拟误差**：上例中估计的 $\hat{\pi}$ 实际是随机的，如果再次独立重复投 N 个点，得到的 $\hat{\pi}$ 和上一次结果会有不同。这是随机模拟方法的特点，即结果具有随机性。因为结果的随机性导致的误差叫做随机模拟误差。

使用随机模拟方法，我们必须了解随机模拟误差的大小，这样我们才能够设计合适的重复试验次数来控制随机模拟误差。比如，这个例子中 ξ 服从 $\mathrm{B}(N,\frac{\pi}{4})$ 分布，有

$$\mathrm{Var}(\hat{\pi})=\frac{\pi(4-\pi)}{16N},$$

由中心极限定理，$\hat{\pi}$ 近似服从 $\mathrm{N}(\pi,\frac{\pi(4-\pi)}{16N})$ 分布，所以随机模拟误差的幅度大约在 $\pm2\sqrt{\frac{\pi(4-\pi)}{16N}}$ (随机模拟误差 95% 以上落入此区间)。

这样的随机误差幅度也是随机模拟误差的典型情况。一般地，假设随机变量 X 期望为 θ，方差为 σ^2。产生随机变量 X 的 N 个独立同分布随机数 $X_i, i=1,2,\ldots,N$, 用样本平均值 $\bar{X}_N=\frac{1}{N}\sum_{i=1}^{N}X_i$ 估计 θ，由中心极限定理可知 $\bar{X}_N$ 近似服从 $\mathrm{N}(\theta,\sigma^2/N)$，于是随机模拟误差幅度为 $O_p(\frac{\sigma}{\sqrt{N}})$ $(N\to\infty)$。为了估计 $\bar{X}_N$ 的渐近标准差 $\sigma/\sqrt{N}$, 可以用样本方差 $S_N^2=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X}_N)^2$ 代替 σ^2 进行计算。为了控制估计的标准差小于 σ_0, 可以先取较小的 N_0 抽取 N_0 个样本值计算出 $S_{N_0}^2$，用 $S_{N_0}^2$ 估计 σ^2，然后求需要的 N 的大小:

$$\frac{S_{N_0}}{\sqrt{N}}<\sigma_0,\quad N>\frac{S_{N_0}^2}{\sigma_0^2}.$$

用 $\bar{X}_N$ 估计 $\theta=EX$ 时，也可以利用中心极限定理计算 θ 的近似 95% 置信区间：

$$\bar{X}_N\pm2S_N/\sqrt{N}. \tag{3.1}$$

随机模拟方法虽然避免了复杂的理论分析，但是其结果具有随机性，精度很难提高：为了增加一位小数点的精度，即误差减小到原来的 $\frac{1}{10}$，重复试验次数需要增加到原来的 100 倍。随机模拟方法有如下特点:

- 应用面广，适应性强。只要问题能够清楚地描述出来，就可以编写模拟程序产生大量数据，通过分析模拟数据解决问题。
- 算法简单, 容易改变条件, 但计算量大。

- 结果具有随机性，精度较低 (一般为 $O(\frac{1}{\sqrt{N}})$ 级)。
- 模拟结果的收敛服从概率论规律。
- 对维数增加不敏感。在计算定积分时，如果使用传统数值算法，维数增加会造成计算时间指数增加，但是如果使用随机模拟方法计算，则维数增加仅仅造成不多的影响。

随机模拟用在科学研究中，常常作为探索性试验来使用。假设科学家有了一个新的模型或技术的想法，但是不知道它的效果怎样所以还没有对其进行深入的理论分析，就可以用随机模拟方法大量地重复生成模拟数据，根据多次重复的总体效果来判断这种模型或技术的性能。如果模拟获得了好的结果，再进行深入理论分析对模型进行完善；如果模拟发现了这个模型的缺点，可以进行有针对性的修改，或者考虑转而其它解决办法。

随机模拟在科学研究中的另一种作用是说明新的模型或技术的有效性。在公开发表的统计学论文中，已经有一半以上的文章包括随机模拟结果（也叫数值结果），用来辅助说明自己提出的模型或方法的有效性。有时因为对模型或方法很难进行彻底的理论分析，仅仅使用大量的随机模拟结果来说明模型或方法的有效性。当然，因为模型都是有可变参数的，随机模拟只能针对某些参数组合给出结果，所以，一般认为仅有模拟结果而没有理论分析结果的研究论文是不全面的。§3.5给出一个用随机模拟说明统计技术优良性的例子。

除了以上应用，随机模拟还是许多新的统计方法的主要工具，例如，蒙特卡洛检验，bootstrap 置信区间和 bootstrap 偏差修正，MCMC。利用大量计算机计算（包括随机模拟）来进行统计推断的统计学分支叫做“计算统计”（computational statistics），在本章后面各节将介绍随机模拟的一些应用和技巧。

3.2 随机模拟积分

某些非随机的问题也可以通过概率模型引入随机变量，化为求随机模型的未知参数的问题。§3.1中用随机投点法估计 π 就是这样的一个例子。

随机模拟解决非随机问题的典型代表是计算定积分。通过对随机模拟定积分的讨论，可以展示随机模拟中大部分的问题和技巧。随机模拟积分也称为蒙特卡洛 (Monte Carlo) 积分 (简称 MC 积分)。实际上，统计中最常见的计算问题就是积分和最优化问题。

3.2.1 随机投点法

设函数 $h(x)$ 在有限区间 $[a,b]$ 上定义且有界，不妨设 $0 \le h(x) \le M$。要计算 $I = \int_a^b h(x)dx$，相当于计算曲线下的区域 $D = \{(x,y) : 0 \le y \le h(x), x \in C = [a,b]\}$ 的面积。为此

在 $G=[a,b]\times(0,M)$ 上均匀抽样 N 次，得随机点 $Z_1,Z_2\ \ldots,Z_N$，$Z_i=(X_i,Y_i),i=1,2,\ldots,N$。令

$$\xi_i=\begin{cases}1, & Z_i\in D\\ 0, & \text{其它}\end{cases},\quad i=1,\ldots,N$$

则 $\{\xi_i\}$ 是独立重复试验结果，$\{\xi_i\}$ 独立同 b(1,p) 分布，

$$p=P(Z_i\in D)=V(D)/V(G)=I/[M(b-a)] \tag{3.2}$$

其中 $V(\cdot)$ 表示区域面积。

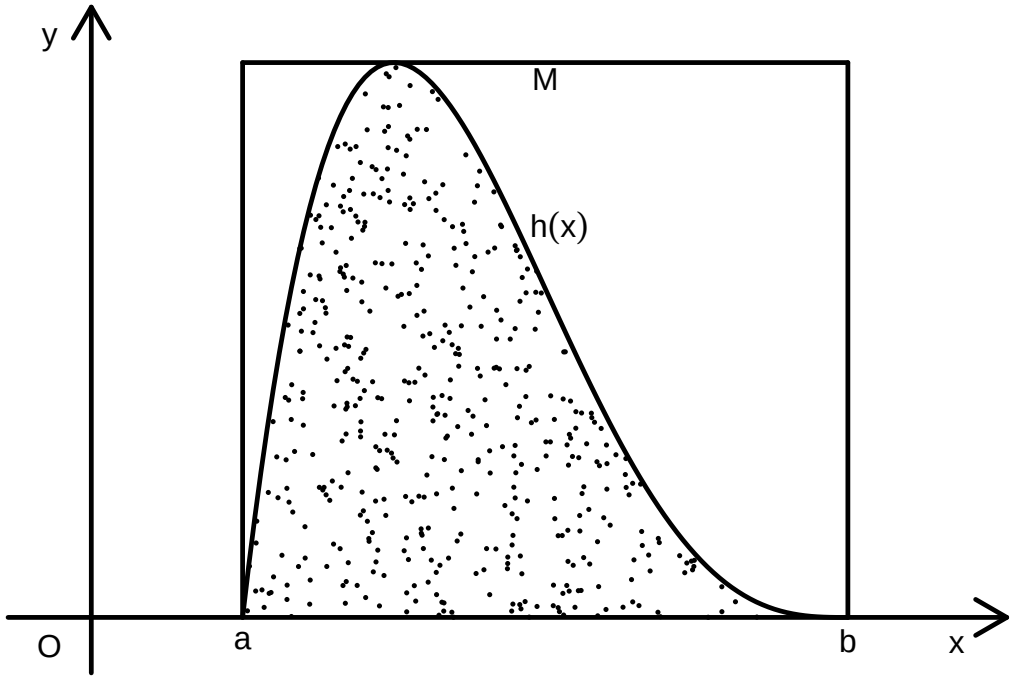

图 3.1: 随机投点法积分示意图

从模拟产生的随机样本 $Z_1,Z_2\ \ldots,Z_N$，可以用这 N 个点中落入曲线下方区域 D 的百分比 $\hat{p}$ 来估计 (3.2)中的概率 p(见图3.1)，然后由 $I=pM(b-a)$ 得到定积分 I 的近似值

$$\hat{I}=\hat{p}M(b-a) \tag{3.3}$$

这种方法叫做**随机投点法**。这样计算的定积分有随机性，误差中包含了随机模拟误差。

由强大数律可知

$$\begin{aligned}\hat{p} =& \frac{\sum \xi_i}{N} \to p, \text{ a.s. } (N \to \infty)\\ \hat{I} =& \hat{p} M(b-a) \to pM(b-a) = I, \text{ a.s. } (N \to \infty)\end{aligned}$$

即 $N \to \infty$ 时精度可以无限地提高（当然，在计算机中要受到数值精度的限制）。

那么，提高精度需要多大的代价呢？由中心极限定理可知

$$\sqrt{N}(\hat{p}-p)/\sqrt{p(1-p)} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1), \ (N \to \infty),$$

从而

$$\sqrt{N}(\hat{I}-I) = M(b-a)(\hat{p}-p) \xrightarrow{\mathrm{d}} \mathrm{N}\big(0, [M(b-a)]^2 p(1-p)\big) \tag{3.4}$$

当 N 很大时 $\hat{I}$ 近似服从 $\mathrm{N}(I, [M(b-a)]^2 p(1-p)/N)$ 分布，称此近似分布的方差 $[M(b-a)]^2 p(1-p)/N$ 为 $\hat{I}$ 的**渐近方差**。计算渐近方差可以用 $\hat{p}$ 代替 p 估计为 $[M(b-a)]^2 \hat{p}(1-\hat{p})/N$。(3.4)说明 $\hat{I}$ 的误差为 $O_p(\frac{1}{\sqrt{N}})$, 这样，计算 $\hat{I}$ 的精度每增加一位小数，计算量需要增加 100 倍。随机模拟积分一般都服从这样的规律。

3.2.2 平均值法

为了计算 $I = \int_a^b h(x)\,dx$，上面用了类似于 §2.2.4的舍选法的做法，在非随机问题中引入随机性时用了二维均匀分布和两点分布，靠求两点分布概率来估计积分 I。随机投点法容易理解，但是效率较低，另一种效率更高的方法是利用期望值的估计。取 $U \sim \mathrm{U}(a,b)$，则

$$\begin{aligned}E[h(U)] =& \int_a^b h(u)\frac{1}{b-a}du = \frac{I}{b-a}\\ I =& (b-a)\cdot Eh(U)\end{aligned}$$

若取 $\{U_i, i=1,\ldots,N\}$ 独立同 $\mathrm{U}(a,b)$ 分布, 则 $Y_i = h(U_i), i=1,2,\ldots,N$ 是 iid 随机变量列, 由强大数律，

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} h(U_i) \to Eh(U) = \frac{I}{b-a}, \quad \text{a.s. } (N \to \infty)$$

于是

$$\tilde{I} = \frac{b-a}{N}\sum_{i=1}^{N} h(U_i) \tag{3.5}$$

是 I 的强相合估计。称这样计算定积分 I 的方法为**平均值法**。由中心极限定理有

$$\sqrt{N}(\tilde{I}-I) \xrightarrow{\mathrm{d}} \mathrm{N}\big(0,(b-a)^2\mathrm{Var}(h(U))\big)$$

其中

$$\mathrm{Var}[h(U)] = \int_a^b \big[h(u)-Eh(U)\big]^2 \frac{1}{b-a}du \tag{3.6}$$

仅与 h 有关，仍有 $\tilde{I}-I=O_p(\frac{1}{\sqrt{N}})$, 但是(3.5)的渐近方差小于(3.3)的渐近方差 (见 §3.2.3)。$\mathrm{Var}[h(U)]$ 可以用模拟样本 $\{Y_i=h(U_i)\}$ 估计为

$$\mathrm{Var}(h(U)) \approx \frac{1}{N}\sum_{i=1}^{N}(Y_i-\bar{Y})^2. \tag{3.7}$$

如果定积分区间是无穷区间，比如 $\int_0^\infty h(x)\,dx$，为了使用均匀分布随机数以及平均值法计算积分可以做积分变换，使积分区间变成有限区间。例如，作变换 $t=1/(x+1)$, 则

$$\int_0^\infty h(x)\,dx = \int_0^1 h(\frac{1}{t}-1)\frac{1}{t^2}\,dt.$$

从平均值法看出，定积分问题 $\int_a^b h(x)\,dx$ 等价于求 $Eh(U)$，其中 $U\sim \mathrm{U}(a,b)$。所以这一节讨论的方法也是用来求随机变量函数期望的随机模拟方法。对一般随机变量 X, 其取值范围不必局限于有限区间，为了求 X 的函数 $h(X)$ 的期望 $I=Eh(X)$，对 X 的随机数 $X_i, i=1,2,\ldots,N$, 令 $Y_i=h(X_i)$, 也可以用平均值法 $\hat{I}=\frac{1}{N}\sum_{i=1}^N h(X_i)$ 来估计 $Eh(X)$, $\hat{I}$ 是 $Eh(X)$ 的无偏估计和强相合估计，若 $\mathrm{Var}(h(X))$ 存在，则 $\hat{I}$ 的方差为 $\frac{1}{N}\mathrm{Var}(h(X))$, $\hat{I}$ 有渐近正态分布 $\mathrm{N}(Eh(X),\frac{1}{N}\mathrm{Var}(h(X)))$。设 $\{Y_i\}$ 的样本方差为 S_N^2，可以用 S_N^2/N 估计 $\hat{I}$ 的方差, 用 $\hat{I}\pm 2S_N/\sqrt{N}$ 作为 I 的近似 95% 置信区间。

例 3.2.1. 设 $X\sim \mathrm{N}(0,1)$，求 $I=E|X|^{\frac{3}{2}}$。

解: 作变量替换积分可得

$$\begin{aligned} I =& 2\int_0^\infty x^{\frac{3}{2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}\,dx \quad (\text{令} t=\frac{1}{2}x^2)\\ =& \frac{2^{\frac{3}{4}}\Gamma(\frac{5}{4})}{\sqrt{\pi}} \approx 0.86004 \end{aligned}$$

如果用平均值法估计 I, 取抽样样本量 $N=10000$, 产生标准正态随机数 $X_i, i=1,2,\ldots,n$, 令 $Y_i=|X_i|^{\frac{3}{2}}$, 令 $\hat{I}=\bar{Y}=\frac{1}{N}\sum_{i=1}^N|X_i|^{\frac{3}{2}}$, $S_N^2=\frac{1}{N}\sum_{i=1}^N(|X_i|^{\frac{3}{2}}-\hat{I})^2$, 一次模拟的结果得到 $\hat{I}=0.8658$, $S_N=0.9409$, 渐近标准差为 $S_N/\sqrt{N}=0.00941$，说明 95% 置信水平下

误差界为 ± 0.02，结果只有将近两位小数的精度。为了达到四位小数的精度，需要误差控制在 ± 0.00005 以下, 需要 $2S_N/\sqrt{N} < 0.00005$, 代入得 N 需要超过 90 亿, 可见随机模拟方法提高精度的困难程度。

以上的评估误差的方法使用了渐近方差和渐近正态分布，在有些模拟问题中估计量不一定有渐近方差，或渐近分布不一定准确（比如产生的随机数不独立时）。在计算量允许的情形下，可以用不同的随机数种子重复得到 B 个 $\hat{I}$ 的估计值，用这 B 个 $\hat{I}$ 的样本标准差来估计 $\hat{I}$ 的抽样分布标准差。取 $B = 100$ 的一次计算结果得到的 $\hat{I}$ 的抽样分布标准差为 0.00924，与从一次模拟得到的渐近标准差结果十分接近。也可以使用 §3.6的 bootstrap 方法获取 B 组 bootstrap 样本，得到 B 个 bootstrap 的 $\hat{I}$ 样本值，从中估计抽样分布标准差。bootstrap 方法可以避免重新计算 $h(X_i)$ 的值, 在更复杂的问题中往往 $h(X_i)$ 的计算是速度很慢的，比如说，$h(X_i)$ 本身也需要用随机模拟或者数值优化方法计算。 □

实际上，一维积分用数值方法均匀布点计算一般更有效。比如，令 $x_i = a + \frac{b-a}{N}i$, $i = 0, 1, \ldots, N$, 估计 I 为 (复合梯形求积公式)

$$I_0 = \frac{b-a}{N}\left[\frac{1}{2}h(a) + \sum_{i=1}^{N-1} h(x_i) + \frac{1}{2}h(b)\right] \tag{3.8}$$

则当 $h(x)$ 在 $[a, b]$ 上二阶连续可微时误差 $I_0 - I$ 仅为 $O(\frac{1}{N^2})$ 阶，比随机模拟方法得到的精度要高得多，而且当 $h(x)$ 有更好的光滑性时还可以用更精确的求积公式得到更高精度。所以，被积函数比较光滑的一元定积分问题一般不需要用随机模拟来计算。

3.2.3 高维定积分

上面的两种计算一元函数定积分的方法可以很容易地推广到多元函数定积分，或称高维定积分。设 d 元函数 $h(x_1, x_2, \ldots, x_d)$ 定义于超矩形

$$C = \{(x_1, x_2, \ldots, x_d) : a_i \le x_i \le b_i, i = 1, 2, \ldots, d\}$$

且

$$0 \le h(x_1, \ldots, x_d) \le M, \ \forall x \in C.$$

令

$$\begin{aligned} D =& \{(x_1, x_2, \ldots, x_d, y) : (x_1, x_2, \ldots, x_d) \in C, \ 0 \le y \le h(x_1, x_2, \ldots, x_d)\}, \\ G =& \{(x_1, x_2, \ldots, x_d, y) : (x_1, x_2, \ldots, x_d) \in C, \ 0 \le y \le M\} \end{aligned}$$

为计算 d 维定积分

$$I=\int_{a_d}^{b_d}\cdots\int_{a_2}^{b_2}\int_{a_1}^{b_1}h(x_1,x_2,\ldots,x_d)\,dx_1dx_2\cdots dx_d, \tag{3.9}$$

产生服从 $d+1$ 维空间中的超矩形 G 内的均匀分布的独立抽样 $\boldsymbol{Z}_1,\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_N$, 令

$$\xi_i=\begin{cases}1, & \boldsymbol{Z}_i\in D\\ 0, & \boldsymbol{Z}_i\in G-D\end{cases},\quad i=1,2,\ldots,N$$

则 ξ_i iid b(1,p),

$$p=P(\boldsymbol{Z}_i\in D)=\frac{V(D)}{V(G)}=\frac{I}{MV(C)}=\frac{I}{M\prod_{j=1}^{d}(b_j-a_j)}$$

其中 $V(\cdot)$ 表示区域体积。令 $\hat{p}$ 为 N 个随机点中落入 D 的百分比，则

$$\hat{p}=\frac{\sum\xi_i}{N}\to p,\ \text{a.s.}(N\to\infty),$$

用

$$\hat{I}_1=\hat{p}V(G)=\hat{p}\cdot M\,V(C)=\hat{p}\cdot M\prod_{j=1}^{d}(b_j-a_j) \tag{3.10}$$

估计积分 I，则 $\hat{I}_1$ 是 I 的无偏估计和强相合估计。称用(3.10)计算高维定积分 I 的方法为随机投点法。由中心极限定理知

$$\sqrt{N}(\hat{p}-p)/\sqrt{p(1-p)}\xrightarrow{d}\mathrm{N}(0,1),$$

$$\sqrt{N}(\hat{I}_1-I)\xrightarrow{d}\mathrm{N}\left(0,\left(M\prod_{j=1}^{n}(b_j-a_j)\right)^2p(1-p)\right),$$

$\hat{I}_1$ 的渐近方差为

$$\frac{\left(M\prod_{j=1}^{n}(b_j-a_j)\right)^2p(1-p)}{N} \tag{3.11}$$

所以 $\hat{I}_1$ 的随机误差仍为 $O_p\left(\frac{1}{\sqrt{N}}\right)$, $N\to\infty$ 时的误差阶不受维数 d 的影响，这是随机模拟方法与其它数值计算方法相比一个重大优势。

在计算高维积分(3.9)时，仍可以通过估计 $Eh(\boldsymbol{U})$ 来获得, 其中 $\boldsymbol{U}$ 服从 R^d 中的超矩形 C 上的均匀分布。设 $\boldsymbol{U}_i \sim$ iid U(C)，$i=1,2,\ldots,N$, 则 $h(\boldsymbol{U}_i), i=1,2,\ldots,N$ 是 iid 随机变量列,

$$Eh(\boldsymbol{U}_i)=\int_C h(\boldsymbol{u})\frac{1}{\prod_{j=1}^d(b_j-a_j)}d\boldsymbol{u}=\frac{I}{\prod_{j=1}^d(b_j-a_j)},$$

估计 I 为

$$\hat{I}_2=\prod_{j=1}^d(b_j-a_j)\cdot\frac{1}{N}\sum_{i=1}^N h(U_i), \tag{3.12}$$

称用(3.12)计算高维定积分 I 的方法为平均值法。由强大数律

$$\hat{I}_2\to\prod_{j=1}^d(b_j-a_j)Eh(U)=I,\ \text{a.s.}(N\to\infty),$$

$\hat{I}_2$ 的渐近方差为

$$\frac{(V(C))^2\mathrm{Var}(h(\boldsymbol{U}))}{N}=\frac{\left(\prod_{j=1}^d(b_j-a_j)\right)^2\mathrm{Var}(h(\boldsymbol{U}))}{N}. \tag{3.13}$$

$N\to\infty$ 时的误差阶也不受维数 d 的影响。

我们来比较随机投点法(3.10)与平均值法(3.12)的精度，只要比较其渐近方差。对 $I=\int_C h(\boldsymbol{x})d\boldsymbol{x}$，设 $\hat{I}_1$ 为随机投点法的估计，$\hat{I}_2$ 为平均值法的估计。因设 $0\le h(\boldsymbol{x})\le M$，不妨设 $0\le h(\boldsymbol{x})\le 1$, 取 $h(\boldsymbol{x})$ 的上界 $M=1$。

令 $\boldsymbol{X}_i\sim$ iid U(C), $\eta_i=h(\boldsymbol{X}_i)$, $Y_i\sim$ iid U(0,1) 与 $\{\boldsymbol{X}_i\}$ 独立，

$$\xi_i=\begin{cases}1, & \text{当}Y_i\le h(\boldsymbol{X}_i)\\ 0, & \text{当}Y_i>h(\boldsymbol{X}_i)\end{cases}\quad i=1,2,\ldots,N,$$

这时有

$$\hat{I}_1=V(C)\frac{1}{N}\sum_{i=1}^N\xi_i,\qquad \hat{I}_2=V(C)\frac{1}{N}\sum_{i=1}^N\eta_i$$

$$\mathrm{Var}(\hat{I}_1)=\frac{1}{N}V^2(C)\cdot\frac{I}{V(C)}\left(1-\frac{I}{V(C)}\right)$$

$$\mathrm{Var}(\hat{I}_2)=\frac{1}{N}V^2(C)\cdot\left(\frac{1}{V(C)}\int_C h^2(\boldsymbol{x})d\boldsymbol{x}-\left(\frac{I}{V(C)}\right)^2\right)$$

$$\mathrm{Var}(\hat{I}_1)-\mathrm{Var}(\hat{I}_2)=\frac{V(C)}{N}\int_C\left\{h(\boldsymbol{x})-h^2(\boldsymbol{x})\right\}d\boldsymbol{x}\ge 0$$

可见平均值法精度更高。事实上，随机投点法多用了随机数 Y_i，必然会增加抽样随机误差。

在计算高维积分(3.10)时，如果用网格法作数值积分，把超矩形 $C=[a_1,b_1]\times[a_2,b_2]\times\cdots\times[a_d,b_d]$ 的每个边分成 n 个格子，就需要 $N=n^d$ 个格子点，如果用每个小超矩形的中心作为代表点，可以达到 $O(n^{-2})$ 精度，即 $O(N^{-2/d})$，当维数增加时为了提高一倍精度需要 $2^{d/2}$ 倍的代表点。比如 $d=8$，精度只有 $O(N^{-1/4})$。高维的问题当维数增加时计算量会迅猛增加，以至于无法计算，这个问题称为**维数诅咒**。如果用 Monte Carlo 积分，则精度为 $O_p(N^{-1/2})$，与 d 关系不大。所以 Monte Carlo 方法在高维积分中有重要应用。为了提高积分计算精度，需要减小 $O_{\mathrm{p}}(N^{-1/2})$ 中的常数项，即减小 $\hat{I}$ 的渐近方差。

例 3.2.2. 考虑

$$f(x_1,x_2,\ldots,x_d)=x_1^2x_2^2\cdots x_d^2,\ 0\le x_1\le 1,0\le x_2\le 1,\ldots,0\le x_d\le 1$$

的积分

$$I=\int_0^1\cdots\int_0^1\int_0^1 x_1^2x_2^2\cdots x_d^2\,dx_1dx_2\cdots dx_d$$

当然，这个积分是有精确解 $I=(1/3)^d$ 的，对估计 I 的结果我们可以直接计算误差。以 $d=8$ 为例比较以下三种方法的精度，这时真值 $I=(1/3)^8\approx 1.524\times 10^{-4}$。

用 n 网格点法，$N=n^d$，公式为

$$I_0=\frac{1}{N}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_d=1}^{n}f(\frac{2i_1-1}{2n},\frac{2i_2-1}{2n},\ldots,\frac{2i_d-1}{2n})$$

误差绝对值为 $e_0=|I_0-I|$。如果取 $n=5,d=8$，需要计算 $N=390625$ 个点的函数值，计算量相当大。

用随机投点法求 I，先在 $(0,1)^d\times(0,1)$ 均匀抽样 $(\xi_i^{(1)},\xi_i^{(2)},\ldots,\xi_i^{(d)},y_i),i=1,\ldots,N$，令 $\hat{I}_1$ 为 $y_i\le f(\xi_i^{(1)},\xi_i^{(2)},\ldots,\xi_i^{(d)})$ 成立的百分比。因为 $\hat{I}_1$ 是随机的，误差绝对值 $|\hat{I}_1-I|$ 也是随机的，所以我们重复试验 B 次，计算 B 次的误差绝对值的平均值 e_1, 作为 $E|\hat{I}_1-I|$ 的估计值。取 B 多大合适呢？因为计算量很大，先取了 $B_1=10$，用得到的 B_1 个 $|\hat{I}_1-I|$ 样本标准差 s_1 可以估计 B 个 $|\hat{I}_1-I|$ 的样本平均值的标准差为 $\mathrm{SE}_1=s_1/\sqrt{B}$，发现 $B=1000$ 时可以控制在相对误差 2% 以下。

用平均值方法，公式为

$$\hat{I}_2=\frac{1}{N}\sum_{i=1}^{N}f(\xi_i^{(1)},\xi_i^{(2)},\cdots,\xi_i^{(d)})$$

其中 $\xi_i^{(j)},i=1,\ldots,N,j=1,\ldots,d$ 独立同 U(0,1) 分布。类似地，重复 $B=1000$ 次，计算 B 次的误差绝对值 $|\hat{I}_2-I|$ 的平均值 e_2，作为 $E|\hat{I}_2-I|$ 的估计。

最后，三种方法计算量相同（都计算 $N = 5^8$ 次函数值）的情况下，得到网格点法的误差绝对值为 $e_0 = 1.2 \times 10^{-5}$, 随机投点法的误差绝对值平均值为 $e_1 = 1.6 \times 10^{-5}$, 平均值法的误差绝对值平均值为 $e_2 = 0.20 \times 10^{-5}$, 此问题结果中平均值法比其它两种方法精度高了至少 5 倍。

3.2.4 重要抽样法

$I = \int_C h(\boldsymbol{x})\, d\boldsymbol{x}$ 中积分区域 C 可能是任意形状的, 也可能无界，$h(\boldsymbol{x})$ 在 C 内各处的取值大小差异可能很大，使得直接用平均值法估计 I 时，很多样本点处于 $|h(\boldsymbol{x})|$ 接近于零的地方，造成浪费，另外使得 $\hat{I}_2$ 的渐近方差 (见(3.13)) 中的 $\mathrm{Var}(h(\boldsymbol{U}))$ 很大 ($\boldsymbol{U} \sim \mathrm{U}(C)$)。为此，考虑用非均匀抽样: $|h(\boldsymbol{x})|$ 大的地方密集投点，$|h(\boldsymbol{x})|$ 小的地方稀疏投点。这样可以有效利用样本。设 $g(\boldsymbol{x}), \boldsymbol{x} \in C$ 是一个密度，其形状与 $|h(\boldsymbol{x})|$ 相近，且当 $g(\boldsymbol{x}) = 0$ 时 $h(\boldsymbol{x}) = 0$，当 $\|\boldsymbol{x}\| \to \infty$ 时 $h(\boldsymbol{x}) = o(g(\boldsymbol{x}))$。称 $g(\boldsymbol{x})$ 为**试投密度**或重要抽样密度。

设 $\boldsymbol{X}_i$ iid $\sim g(\boldsymbol{x})$, $i = 1, 2, \ldots, N$。令

$$\eta_i = \frac{h(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)},\ i = 1, 2, \ldots, N$$

则

$$E\eta_1 = \int_C \frac{h(\boldsymbol{x})}{g(\boldsymbol{x})} g(\boldsymbol{x})\, d\boldsymbol{x} = \int_C h(\boldsymbol{x})\, d\boldsymbol{x} = I$$

因此可以用 $\{\eta_i, i = 1, 2, \ldots, N\}$ 的样本平均值来估计 I，即

$$\hat{I}_3 = \bar{\eta} = \frac{1}{N} \sum_{i=1}^{N} \frac{h(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)}. \tag{3.14}$$

$\hat{I}_3$ 是 I 的无偏估计和强相合估计。$\hat{I}_3$ 的渐近方差为

$$\mathrm{Var}(\hat{I}_3) = \mathrm{Var}\left(\frac{h(\boldsymbol{X})}{g(\boldsymbol{X})}\right) \frac{1}{N} = O(\frac{1}{N}), \tag{3.15}$$

当 $g(\boldsymbol{x})$ 和 $|h(\boldsymbol{x})|$ 形状很接近时 $\frac{|h(\boldsymbol{x})|}{g(\boldsymbol{x})}$ 近似为常数，方差 $\mathrm{Var}(\hat{I}_3)$ 很小, 这时 $\hat{I}_3$ 的随机误差可以很小。用(3.14)估计 I 与 §2.2.4的舍选法 II 有类似的想法，这种方法叫做**重要抽样法** (importance sampling)，是随机模拟的重要方法。

为什么当 $g(\boldsymbol{x})$ 和 $|h(\boldsymbol{x})|$ 形状很接近时 $\mathrm{Var}(\hat{I}_3)$ 很小？事实上，

$$\begin{aligned}\mathrm{Var}\left(\frac{h(\boldsymbol{X})}{g(\boldsymbol{X})}\right) =& E\left(\frac{h^2(\boldsymbol{X})}{g^2(\boldsymbol{X})}\right) - \left(E\frac{h(\boldsymbol{X})}{g(\boldsymbol{X})}\right) = E\left(\frac{h^2(\boldsymbol{X})}{g^2(\boldsymbol{X})}\right) - \left(\int_C h(\boldsymbol{x})\,d\boldsymbol{x}\right)^2 \\ \geq& \left(E\frac{|h(\boldsymbol{X})|}{g(\boldsymbol{X})}\right)^2 - I^2 \quad (\text{由 Jensen 不等式}) \qquad (3.16)\\ =& \left(\int_C |h(\boldsymbol{x})|\,d\boldsymbol{x}\right)^2 - I^2, \qquad (3.17)\end{aligned}$$

当且仅当 $\frac{|h(\boldsymbol{X})|}{g(\boldsymbol{X})}$ 为常数时不等式(3.16)中的等号成立，即 $g(\boldsymbol{x}) = \frac{1}{\int_C |h(\boldsymbol{t})|\,d\boldsymbol{t}}|h(\boldsymbol{x})|, \boldsymbol{x} \in C$ 时 $\mathrm{Var}(\hat{I}_3)$ 达到最小。

上述求积分的问题也可以表述为求随机变量函数的期望问题。设 $\boldsymbol{Y} \sim f(\boldsymbol{y})$，要求 $\boldsymbol{Y}$ 的函数 $h(\boldsymbol{Y})$ 的期望值 $Eh(\boldsymbol{Y}) = \int h(\boldsymbol{y})f(\boldsymbol{y})\,d\boldsymbol{y}$, 可以抽取 $\boldsymbol{Y}$ 的随机数 $\boldsymbol{Y}_i, i = 1, 2, \ldots, N$, 然后用平均值法 $\frac{1}{N}\sum_{i=1}^N h(\boldsymbol{Y}_i)$ 估计 $Eh(\boldsymbol{Y})$。但是，如果很难生成 $\boldsymbol{Y}$ 的随机数，或者 $\boldsymbol{Y}$ 的分布集中于 $h(\boldsymbol{y})$ 接近于零的位置以至于积分效率很低，可以找一个试投密度 $g(\boldsymbol{x})$，从 $g(\boldsymbol{x})$ 产生随机数 $\boldsymbol{X}_i, i = 1, 2, \ldots, N$, 用如下重要抽样法

$$\hat{I}_{3.1} = \frac{1}{N}\sum_{i=1}^N h(\boldsymbol{X}_i)\frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)} \qquad (3.18)$$

估计 $Eh(\boldsymbol{Y})$, 易见 $E\hat{I}_{3.1} = Eh(\boldsymbol{Y})$。称 $W_i = \frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)}$ 为**重要性权重**, $Eh(\boldsymbol{Y})$ 可以用重要性权重估计为

$$\hat{I}_{3.1} = \frac{1}{N}\sum_{i=1}^N W_i h(\boldsymbol{X}_i). \qquad (3.19)$$

选取试投密度 $g(\boldsymbol{x})$ 时，要求当 $h(\boldsymbol{x})f(\boldsymbol{x}) \neq 0$ 时 $g(\boldsymbol{x}) \neq 0$, 当 $\boldsymbol{x}$ 趋于无穷时 $h(\boldsymbol{x})f(\boldsymbol{x}) = o(g(\boldsymbol{x}))$（$g(\boldsymbol{x})$ 相对厚尾）, 一般还要求 $g(\boldsymbol{x})$ 的形状与 $|h(\boldsymbol{x})|f(\boldsymbol{x})$ 形状接近。如果 $g(\boldsymbol{x})$ 的相对厚尾性难以确定，可以使用如下保险的试投密度

$$\tilde{g}(\boldsymbol{x}) = \rho g(\boldsymbol{x}) + (1-\rho)r(\boldsymbol{x}), \qquad (3.20)$$

其中 ρ 接近于 1，$r(\boldsymbol{x})$ 是柯西分布或 Pareto 分布这样的重尾分布。要生成 N 个 $\tilde{g}(\boldsymbol{x})$ 的随机数，只要生成 $N\rho$ 个 $g(\boldsymbol{x})$ 的随机数和 $N(1-\rho)$ 个 $r(\boldsymbol{x})$ 的随机数。

选取不适当的试投密度会把绝大多数样本点投到了对计算积分不重要的位置，使得样本点中只有极少数点是真正有作用的。在多维问题中合适的试投密度尤其难找，经常需要反复试验。

有时 $\boldsymbol{Y} \sim f(\cdot)$ 不仅很难直接抽样，而且 $f(\cdot)$ 本身未知，只能确定到差一个常数倍的 $\tilde{f}(\boldsymbol{x}) = cf(\boldsymbol{x})$, 常数 c 未知，为了求常数 c 需要计算 $c = \int \tilde{f}(\boldsymbol{y})\,d\boldsymbol{y}$，计算 c 一般很困难。这时，定义重要性权重为 $W_i = \frac{\tilde{f}(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)}$, 公式(3.19)可以改成

$$\hat{I}_4 = \frac{\sum_{i=1}^N W_i h(\boldsymbol{X}_i)}{\sum_{i=1}^N W_i} \tag{3.21}$$

这称为**标准化重要抽样法**。对(3.21)的分子和分母都除以 cN 后分子 a.s. 收敛到 $Eh(\boldsymbol{Y})$, 分母 a.s. 收敛到 1, 所以(3.21)是 $Eh(\boldsymbol{Y})$ 的强相合估计，但不是无偏的。标准化重要抽样估计往往比无偏估计 $\hat{I}_{3.1}$ 有更小的均方误差。关于标准化重要抽样法的渐近方差的讨论参见 Liu(2001)[28]§2.5.3。

如果需要对多个不同的函数 $h(\cdot)$ 计算 $Eh(\boldsymbol{Y})$, 则选取试抽样密度 $g(\boldsymbol{x})$ 时应使得 $g(\boldsymbol{x})$ 尽可能与 $\boldsymbol{Y}$ 的密度 $f(\boldsymbol{x})$ 形状接近，这样权重 $W_i = \frac{\tilde{f}(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)}$ 的分布不过于偏斜，不至于出现绝大部分权重集中于少数样本点的情形。抽样值 $\boldsymbol{X}_i$ 与权重 W_i 一起可以看作是分布 $f(\cdot)$ 的某种抽样。

定义 (**适当加权抽样**) 随机变量序列 $\{(\boldsymbol{X}_i, W_i),\, i = 1, 2, \ldots, N\}$ 称为关于密度 $f(\cdot)$ 的**适当加权抽样**, 如果对于任何平方可积函数 $h(\cdot)$ 都有

$$E[h(X_i)W_i] = cE_f[h(X)] = c\int h(\boldsymbol{x})f(\boldsymbol{x})\,d\boldsymbol{x},\ \ i = 1, 2, \ldots, N,$$

其中 c 是归一化常数。

设随机变量 $(\boldsymbol{X}, W)$ 联合密度为 $g(\boldsymbol{x}, w)$，则 $(\boldsymbol{X}, W)$ 的样本为密度 $f(\cdot)$ 的适当加权抽样的充分必要条件是

$$E_g(W|\boldsymbol{x}) = E_g(W)\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})},\ \ \forall \boldsymbol{x},$$

其中 $E_g(W)$ 是关于 W 的边缘密度的期望，$E_g(W|\boldsymbol{x})$ 是在 $(\boldsymbol{X}, W)$ 的联合密度下条件期望 $E(W|\boldsymbol{X})$ 在 $\boldsymbol{X} = \boldsymbol{x}$ 处的值。

在重要抽样法和标准化重要抽样法的实际应用中，好的试抽样分布很难获得，所以权重 $\{W_i = f(\boldsymbol{X}_i)/g(\boldsymbol{X}_i)\}$ 经常会差别很大，使得抽样样本主要集中在少数几个权重最大的样本点上。为此，可以舍弃权重太小的样本点，重新抽样替换这样的样本点。这种方法称为**舍选控制** (Rejection Control)，描述如下。

首先，需要选定权重的一个阈值 c, 然后对每个样本点 $\boldsymbol{X}_i$, 若 $W_i \geq c$, 则接受 $\boldsymbol{X}_i$; 若 $W_i < c$, 则以概率 W_i/c 接受 $\boldsymbol{X}_i$，否则舍弃 $\boldsymbol{X}_i$，重新抽取。最后，所有样本点的权重调整为新的

$$W_i^* = p_c\frac{W_i}{\min\{1,\ W_i/c\}} = p_c\max\{W_i, c\}, \tag{3.22}$$

其中

$$p_c = \int \min\left\{1, \frac{f(\boldsymbol{x})}{cg(\boldsymbol{x})}\right\} g(\boldsymbol{x})\, d\boldsymbol{x} = \int \min\left\{g(\boldsymbol{x}), \frac{f(\boldsymbol{x})}{c}\right\} d\boldsymbol{x} \tag{3.23}$$

是归一化常数，如果使用标准化重要抽样法，p_c 可以省略，否则，p_c 可以估计为

$$\hat{p}_c = \frac{1}{N}\sum_{i=1}^{N} \min\left\{1,\ \frac{W_i}{c}\right\}. \tag{3.24}$$

这样得到的 $\{(\boldsymbol{X}_i, W_i^*),\ i = 1, \ldots, N\}$ 是关于 $f(\cdot)$ 适当加权的, 被接受的 $\boldsymbol{X}_i$ 的分布密度为

$$g^*(\boldsymbol{x}) = \frac{1}{p_c} \min\left\{g(\boldsymbol{x}), \frac{f(\boldsymbol{x})}{c}\right\}. \tag{3.25}$$

阈值 c 在实际中可以从权重 $\{W_i\}$ 选取，比如，取为 $\{W_i\}$ 的某个分位数。

例 3.2.3. 用 MC 积分法计算 $I = \int_0^1 e^x\, dx = e - 1 \approx 1.718$。对被积函数 $h(x) = e^x$ 做泰勒展开得

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots$$

取

$$g(x) = c(1 + x) = \frac{2}{3}(1 + x)$$

要产生 $g(x)$ 的随机数可以用逆变换法，密度 $g(x)$ 的分布函数 $G(x)$ 的反函数为

$$G^{-1}(y) = \sqrt{1 + 3y} - 1,\ 0 < y < 1$$

因此，取 U_i iid U(0,1), 令 $X_i = \sqrt{1 + 3U_i} - 1$, $i = 1, 2, \ldots, N$, 则重要抽样法的积分公式为

$$\hat{I}_3 = \frac{1}{N}\sum_{i=1}^{N} \frac{e^{X_i}}{\frac{2}{3}(1 + X_i)}$$

渐近方差为

$$\mathrm{Var}(\hat{I}_3) = \frac{1}{N}\left(\frac{3}{2}\int_0^1 \frac{e^{2x}}{1 + x} dx - I^2\right) \approx 0.02691/N.$$

如果用平均值法，估计公式为

$$\hat{I}_2 = \frac{1}{N}\sum_{i=1}^{N} e^{U_i},$$

渐近方差为

$$\mathrm{Var}(\hat{I}_2)=\frac{1}{N}\left(\int_0^1 e^{2x}\,dx-I^2\right)\approx 0.2420/N\approx 9.0\times\mathrm{Var}(\hat{I}_3)$$

是重要抽样法方差的 9 倍。

如果用随机投点法，$h(x)=e^x\le e(0<x<1)$, 取上界 $M=e$, 向 $[0,1]\times[0,M]$ 随机投点，落到 $f(x)$ 下方的概率为

$$p=I/(M(b-a))=(e-1)/e,$$

设投 N 点落到 $h(x)$ 下方的频率为 $\hat{p}$, 用随机投点法估计 I 的公式为

$$\hat{I}_1=\hat{p}\cdot M(b-a)=e\hat{p},$$

渐近方差为

$$\mathrm{Var}(\hat{I}_1)=e^2p(1-p)/N=(e-1)/N\approx 1.718/N\approx 7.1\times\mathrm{Var}(\hat{I}_2)\approx 64.8\times\mathrm{Var}(\hat{I}_3)$$

可见选择合适的抽样算法对减少计算量、提高精度是十分重要的。 □

例 3.2.4. 设二元函数 $f(x,y)$ 定义如下

$$f(x,y)=\exp\{-45(x+0.4)^2-60(y-0.5)^2\}+0.5\exp\{-90(x-0.5)^2-45(y+0.1)^4\}$$

求如下二重定积分

$$I=\int_{-1}^1\int_{-1}^1 f(x,y)\,dx\,dy$$

$f(x,y)$ 有两个分别以 $(-0.4,0.5)$ 和 $(0.5,-0.1)$ 为中心的峰，对积分有贡献的区域主要集中在 $(-0.4,0.5)$ 和 $(0.5,-0.1)$ 附近，在其他地方函数值很小，对积分贡献很小。

用平均值法(3.12)，取点数 $N=10000$ 时 $\hat{I}_2$ 的一个估计值为 $\hat{I}_2=0.132$, 从这一次模拟估计的 $\hat{I}_2$ 的渐近方差为 1.86×10^{-5}。重复模拟 $B=100$ 次，得到这 100 个 $\hat{I}_2$ 的平均值为 0.1256, 样本方差为 1.95×10^{-5}。

用重要抽样法。取试投密度为

$$\begin{aligned}g(x,y)\propto&\tilde{g}(x,y)\\&=\exp\{-45(x+0.4)^2-60(y-0.5)^2\}+0.5\exp\{-90(x-0.5)^2-10(y+0.1)^2\},\\&\quad-\infty<x<\infty,-\infty<y<\infty,\end{aligned}$$

这样抽取到 $[-1,1]\times[-1,1]$ 范围外的点对积分没有贡献，因为构成 $g(x,y)$ 的两个密度都很集中，所以效率损失不大。需要求使得 $\tilde{g}(x,y)$ 化为密度的比例常数。记 $\mathrm{N}(\mu,\sigma^2)$ 的分布密度为 $f(x;\mu,\sigma^2)$，对 $\tilde{g}(x,y)$ 积分得

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\tilde{g}(x,y) =&\sqrt{2\pi/90}\int_{-\infty}^{\infty}f(x;-0.4,90^{-1})dx\cdot\sqrt{2\pi/120}\int_{-\infty}^{\infty}f(y;0.5,120^{-1})dy\\ &+0.5\sqrt{2\pi/180}\int_{-\infty}^{\infty}f(x;0.5,180^{-1})dx\cdot\sqrt{2\pi/20}\int_{-\infty}^{\infty}f(y;-0.1,20^{-1})dy\\ =&\sqrt{2\pi/90}\sqrt{2\pi/120}+0.5\sqrt{2\pi/180}\sqrt{2\pi/20}\\ \approx&0.1128199\end{aligned}$$

于是令

$$\begin{aligned}g(x,y) =&\tilde{g}(x,y)/0.1128199\\ =&0.5358984f(x;-0.4,90^{-1})f(y;0.5,120^{-1})\\ &+0.4641016f(x;0.5,180^{-1})f(y;-0.1,20^{-1}),\\ &-\infty<x<\infty,-\infty<y<\infty,\end{aligned}$$

用复合抽样法对 $g(x,y)$ 抽样，$N=10000$ 时一次模拟得到的 $\hat{I}_3=0.126$, 从这一次模拟估计的 $\hat{I}_3$ 的渐近方差为 6.84×10^{-8}, 重复模拟 $B=100$ 次，则 100 个 $\hat{I}_3$ 的平均值为 0.1258, 样本方差为 5.37×10^{-8}。$\hat{I}_2$ 的样本方差是 $\hat{I}_3$ 的样本方差的 363 倍，如果要达到相同的估计方差，两种方法的样本量相差三百多倍。 □

例 3.2.5. 标准化的重要抽样法在贝叶斯统计推断中有重要作用。例如，设独立的观测样本 Y_j 服从如下的贝塔—二项分布：

$$f(y_j|K,\eta)=P(Y_j=y_j)=\binom{n_j}{y_j}\frac{B(K\eta+y_j,\ K(1-\eta)+n_j-y_j)}{B(K\eta,\ K(1-\eta))},\ y_j=0,1,\ldots,n_j,\quad(3.26)$$

其中 $B(\cdot,\cdot)$ 是贝塔函数，n_j 为已知的正整数，$K>0$, $0<\eta<1$ 为未知参数。贝塔—二项分布用于描述比二项分布更为分散的随机变量分布。按照贝叶斯统计的做法，假设参数 (K,η) 也是随机变量，具有所谓的“先验分布”，假设 (K,η) 有如下的“无信息”先验分布密度:

$$\pi(K,\eta)\propto\frac{1}{(1+K)^2}\frac{1}{\eta(1-\eta)},\qquad(3.27)$$

则 (K,η) 有如下的“后验密度”：

$$\begin{aligned}\tilde{p}(K,\eta|\boldsymbol{Y}) \propto & \pi(K,\eta)\prod_{j=1}^{n} f(y_j|K,\eta)\\ \propto & \frac{1}{(1+K)^2}\frac{1}{\eta(1-\eta)}\prod_{j=1}^{n}\frac{B(K\eta+y_j,\ K(1-\eta)+n_j-y_j)}{B(K\eta,\ K(1-\eta))}.\end{aligned} \tag{3.28}$$

设要求 $E(\log K|\boldsymbol{Y})=\int_0^\infty \log K\,\tilde{p}(K,\eta|\boldsymbol{Y})\,dK$ 的值。

如果可以从后验密度 $\tilde{p}(K,\eta|\boldsymbol{Y})$ 直接抽样，可以用平均值法估计 $E(\log K|\boldsymbol{Y})$，但从(3.28)来看很难直接抽样。为此，使用标准化的重要抽样法。为了解除 (K,η) 的取值限制，作变换 $\alpha=\log K$, $\beta=\log\frac{\eta}{1-\eta}$, 则 $\alpha,\beta\in(-\infty,\infty)$, 而(3.28)对应的 (α,β) 的后验密度为:

$$p(\alpha,\beta|\boldsymbol{Y})\propto\frac{e^\alpha}{(1+e^\alpha)^2}\prod_{j=1}^{n}\frac{B(\frac{e^\alpha}{1+e^{-\beta}}+y_j,\ \frac{e^\alpha}{1+e^\beta}+n_j-y_j)}{B(\frac{e^\alpha}{1+e^{-\beta}},\ \frac{e^\alpha}{1+e^\beta})}. \tag{3.29}$$

取值无限制的随机变量试抽样密度经常使用自由度较小的 t 分布，比如 t(4) 分布，设 t(4) 分布密度函数为 $g(\cdot)$, 用独立的 t(4) 分布生成 (α,β) 的试抽样样本 $(\alpha_i,\beta_i), i=1,2,\ldots,N$，可以估计 $E(\log K|\boldsymbol{Y})$ 为

$$\hat{\alpha}=\frac{\sum_{i=1}^{N}\alpha_i\frac{p(\alpha_i,\beta_i|\boldsymbol{Y})}{g(\alpha_i)g(\beta_i)}}{\sum_{i=1}^{N}\frac{p(\alpha_i,\beta_i|\boldsymbol{Y})}{g(\alpha_i)g(\beta_i)}}. \tag{3.30}$$

其中的 $p(\alpha_i,\beta_i|\boldsymbol{Y})$ 只要用(3.29)的右侧计算，因为分子和分母的归一化常数可以消掉。 □

3.2.5 分层抽样法

用平均值法计算 $\int_C h(\boldsymbol{x})\,d\boldsymbol{x}$，若 $h(\boldsymbol{x})$ 在 C 内取值变化范围大则估计方差较大。重要抽样法选取了与 $f(x)$ 形状相似但是容易抽样的密度 $g(\boldsymbol{x})$ 作为试投密度，大大提高了精度，但是这样的 $g(\boldsymbol{x})$ 有时难以找到。

如果把 C 上的积分分解为若干个子集上的积分，使得 $h(\boldsymbol{x})$ 在每个子集上变化不大，分别计算各个子集上的积分再求和，可以提高估计精度。这种方法与 §2.2.5的复合抽样法类似，叫做**分层抽样法**。这也是抽样调查中的重要技术。

例 3.2.6. 对函数

$$h(x)=\begin{cases}1+\frac{x}{10}, & 0\le x\le 0.5\\ -1+\frac{x}{10}, & 0.5<x\le 1\end{cases}$$

求定积分

$$I=\int_0^1 h(x)\,dx,$$

可以得 I 的精确值为 $I=0.05$。我们用平均值法和分层抽样法来估计 I 并比较精度。

在 [0,1] 区间随机抽取 N 点用平均值法得 $\hat{I}_2$，其渐近方差为

$$\mathrm{Var}(\hat{I}_2)=\frac{\mathrm{Var}(h(U))}{N}=\frac{143}{150N}\approx\frac{0.9533}{N}.$$

把 I 拆分为 $[0,0.5]$ 和 $[0.5,1]$ 上的积分，即

$$I=a+b=\int_0^{0.5}h(x)\,dx+\int_{0.5}^1 h(x)\,dx,$$

对 a 和 b 分别用平均值法，得

$$\begin{aligned}\hat{a}=&\frac{0.5}{N/2}\sum_{i=1}^{N/2}h(0.5U_i)=\frac{0.5}{N/2}\sum_{i=1}^{N/2}(1+0.05U_i),\\ \hat{b}=&\frac{0.5}{N/2}\sum_{i=(N/2)+1}^{N}h(0.5+0.5U_i)=\frac{0.5}{N/2}\sum_{i=(N/2)+1}^{N}(-1+0.05+0.05U_i),\\ \hat{I}_5=&\hat{a}+\hat{b},\end{aligned}$$

则分层抽样法结果 $\hat{I}_5$ 的渐近方差为

$$\begin{aligned}\mathrm{Var}(\hat{I}_5)=&\mathrm{Var}(\hat{a}+\hat{b})=\mathrm{Var}(\hat{a})+\mathrm{Var}(\hat{b})\\ =&0.25\frac{\mathrm{Var}(1+0.05U)}{N/2}+0.25\frac{\mathrm{Var}(-0.95+0.05U)}{N/2}=\frac{1/4800}{N},\end{aligned}$$

分层后的估计方差远小于不分层的结果, 可以节省样本量约 4500 倍。 □

一般地，设积分 $I=\int_C h(\boldsymbol{x})\,d\boldsymbol{x}$ 可以分解为 m 个不交的子集 C_j 上的积分, 即

$$I=\int_C h(\boldsymbol{x})\,d\boldsymbol{x}=\int_{C_1}h(\boldsymbol{x})\,d\boldsymbol{x}+\int_{C_2}h(\boldsymbol{x})\,d\boldsymbol{x}+\cdots+\int_{C_m}h(\boldsymbol{x})\,d\boldsymbol{x}$$

在 C_j 投 n_j 个随机点 $X_{ji}\sim \mathrm{U}(C_j)$, $i=1,\ldots,n_j$, 则 I 的 m 个部分可以分别用平均值法估计，由此得 I 的分层估计为

$$\hat{I}_5=\sum_{j=1}^{m}\frac{V(C_j)}{n_j}\sum_{i=1}^{n_j}h(X_{ji})$$

记 $\sigma_j^2 = \mathrm{Var}(h(X_{j1}))$, 划分子集时应使每一子集内 $h(\cdot)$ 变化不大，即 σ_j^2 较小。这时

$$\mathrm{Var}(\hat{I}_5) = \sum_{j=1}^{m} \frac{V^2(C_j)\sigma_j^2}{n_j}$$

若 σ_j^2 可估计，应取 n_j 使

$$n_j \propto V(C_j)\sigma_j, \tag{3.31}$$

即

$$n_j = N\frac{V(C_j)\sigma_j}{\sum_{k=1}^{m} V(C_k)\sigma_k},\ j = 1, 2, \ldots, m$$

这样取的样本量 $(n_1, n_2, \ldots, n_m)$ 在所有满足 $n_1 + n_2 + \cdots + n_m = N$ 的取法中使得渐近方差最小 (见习题1)。

在分层抽样法中，划分了子集后，每一子集上的积分也可用重要抽样法计算。

分层抽样法也可以用在求随机变量函数期望的问题中。设 X 为随机变量，要求 X 的函数 $h(X)$ 的数学期望 $\theta = Eh(X)$。假设存在离散型随机变量 Y，$p_j = P(Y = y_j), j = 1, 2, \ldots, m$, 在 $Y = y_j$ 条件下可以从 X 的条件分布抽样，则

$$E[h(X)] = E\{E[h(X)|Y]\} = \sum_{j=1}^{m} E[h(X)|Y = y_j]p_j, \tag{3.32}$$

如果在 $Y = y_j$ 条件下生成 X 的 $N_j = Np_j$ 个抽样值，设为 $X_i^{(j)}, i = 1, 2, \ldots, N_j$, 则可以用 $\frac{1}{N_j}\sum_{i=1}^{N_j} h(X_i^{(j)})$ 估计 $E[h(X)|Y = y_j]$，估计 θ 为

$$\hat{\theta} = \sum_{j=1}^{m} \frac{1}{N_j} \sum_{i=1}^{N_j} h(X_i^{(j)})p_j = \frac{1}{N} \sum_{j=1}^{m} \sum_{i=1}^{Np_j} h(X_i^{(j)}), \tag{3.33}$$

这是 θ 的无偏和强相合估计，且估计方差

$$\begin{aligned}
\mathrm{Var}(\hat{\theta}) =& \frac{1}{N^2} \sum_{j=1}^{m} Np_j \mathrm{Var}[h(X)|Y = y_j] \\
=& \frac{1}{N} \sum_{j=1}^{m} \mathrm{Var}[h(X)|Y = y_j]p_j \qquad (3.34) \\
=& \frac{1}{N} E\{\mathrm{Var}[h(X)|Y]\} \le \frac{1}{N}\mathrm{Var}[h(X)], \qquad (3.35)
\end{aligned}$$

比直接用平均值法估计 $Eh(X)$ 的方差小。这里用到了条件方差的性质

$$\mathrm{Var}(X) = E\left[\mathrm{Var}(X|Y)\right] + \mathrm{Var}\left[E(X|Y)\right] \geq E\left[\mathrm{Var}(X|Y)\right], \tag{3.36}$$

如果 Y 与 X 独立则 $E(X|Y) = EX$, $\mathrm{Var}\left[E(X|Y)\right] = 0$, 这时分层抽样法比平均值法没有改进。从(3.34)可以看出，如果第 j 层样本的函数 $h(X_i^{(j)}), i = 1, 2, \ldots, Np_j$ 的样本方差为 S_j^2, 则 $\mathrm{Var}(\hat{\theta})$ 的一个无偏估计是

$$\widehat{\mathrm{Var}(\hat{\theta})} = \frac{1}{N}\sum_{j=1}^{m} S_j^2 p_j. \tag{3.37}$$

公式(3.33)取第 j 层样本数 $N_j = Np_j$，仅考虑了 Y 的取值分布，而未考虑 $X|Y = y_j$ 的条件分布情况。类似于(3.31)和(3.32)，应该对 $\mathrm{Var}[h(X)|Y = y_j]$ 较大的层取更多的样本。使得估计方差最小的分层样本量分配满足 $N_j \propto p_j\sqrt{\mathrm{Var}[h(X)|Y = y_j]}$, 即

$$N_j = N\frac{p_j\sqrt{\mathrm{Var}[h(X)|Y = y_j]}}{\sum_{k=1}^{m} p_k\sqrt{\mathrm{Var}[h(X)|Y = y_k]}}. \tag{3.38}$$

在 $\mathrm{Var}[h(X)|Y = y_j]$ 未知的时候，可以预先抽取一个小的样本估计 $\mathrm{Var}[h(X)|Y = y_j]$, 然后按估计的最优 N_j 分配各层的样本量。采用(3.38)的分层样本量后，

$$\hat{\theta} = \sum_{j=1}^{m}\frac{1}{N_j}\sum_{i=1}^{N_j} h(X_i^{(j)})p_j, \tag{3.39}$$

$$\mathrm{Var}(\hat{\theta}) = \sum_{j=1}^{m}\frac{p_j^2\mathrm{Var}[h(X)|Y = y_j]}{N_j}, \tag{3.40}$$

于是 $\mathrm{Var}(\hat{\theta})$ 的估计为

$$\widehat{\mathrm{Var}(\hat{\theta})} = \sum_{j=1}^{m}\frac{p_j^2 S_j^2}{N_j}, \tag{3.41}$$

其中 S_j^2 是第 j 层样本函数 $\{h(X_i^{(j)}), i = 1, 2, \ldots, N_j\}$ 的样本方差。

分层抽样法的本质是把 X 的值相近的抽样分入一层，使得同层的 X 条件方差较小，从而减小估计方差。

例 3.2.7. 设 $U \sim \mathrm{U}(0, 1)$, 要估计 $\theta = Eh(U) = \int_0^1 h(x)\,dx$。令 $Y = \mathrm{ceil}(mU)$, 即当且仅当

$\frac{j-1}{m} < U \le \frac{j}{m}$ 时 $Y = j$，$j = 1, 2, \ldots, m$, 可以按照 Y 分层抽样估计 θ:

$$\theta = E[h(U)] = \sum_{j=1}^{m} E[h(U)|Y = j]\, P(Y = j) \tag{3.42}$$

$$= \frac{1}{m} \sum_{j=1}^{m} E[h(U)|Y = j], \tag{3.43}$$

易见 $Y = j$ 条件下 U 服从 $(\frac{j-1}{m}, \frac{j}{m})$ 上的均匀分布，设 $U_1, U_2, \ldots, U_n$ 是 U(0,1) 的独立抽样，则用分层抽样法取每层 $N_j = 1$ 估计 $\theta = Eh(U)$ 为

$$\hat{\theta} = \frac{1}{m} \sum_{j=1}^{m} h\left(\frac{j - 1 + U_j}{m}\right). \tag{3.44}$$

□

3.3 方差缩减方法

随机模拟方法虽然有着适用性广、方法简单的优点，但是又有精度低、计算量大的缺点, 一整套模拟算几天几夜也是常有的事情。如果能成倍地减小随机模拟误差方差，就可以有效地节省随机模拟时间，有些情况下可以把耗时长到不具有可行性的模拟计算（比如几个月）缩短到可行（比如几天）。

前一节的关于定积分计算的重要抽样法、分层抽样法都是降低随机模拟误差方差的重要方法，也可以用在一般的模拟问题中。本节介绍一些其它的方差缩减技巧。我们以随机变量 X 的期望 $\theta = EX$ 的估计为例，目标是降低 θ 的估计量的渐近方差。

3.3.1 控制变量法

设要估计随机变量 X 的期望 $\theta = EX$, 从 X 中抽取 N 个独立样本值 $X_1, X_2, \ldots, X_n$，用样本平均值 $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$ 估计 EX。为了提高精度可以利用辅助信息。设有另外的随机变量 Y 满足

$$EY = 0, \quad \mathrm{Cov}(X, Y) < 0$$

令 $Z = X + Y$, 则

$$E(Z) = \theta, \quad \mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y),$$

只要 $\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)<0$ 则 $\mathrm{Var}(Z)<\mathrm{Var}(X)$，如果有 (X,Y) 成对的抽样 $(X_i,Y_i), i=1,2,\ldots,n$, 令 $Z_i=X_i+Y_i$, 则用 $\bar{Z}$ 来估计 $\theta=EX=EZ$ 的渐近方差就比用 $\bar{X}$ 估计 I 的渐近方差减小了。

为了最好地利用 Y 与 X 的相关性，令

$$Z(b)=X+bY,$$

则

$$\begin{aligned} EZ(b) =& EX=\theta, \\ \mathrm{Var}(Z(b)) =& \mathrm{Var}(X)+2b\mathrm{Cov}(X,Y)+b^2\mathrm{Var}(Y), \end{aligned}$$

求 $\mathrm{Var}(Z(b))$ 关于 b 的最小值点，得

$$b=-\mathrm{Cov}(X,Y)/\mathrm{Var}(Y)=-\rho_{X,Y}\sqrt{\mathrm{Var}(X)/\mathrm{Var}(Y)},$$

这时

$$\mathrm{Var}(Z(b))=(1-\rho_{X,Y}^2)\mathrm{Var}(X)\le\mathrm{Var}(X),$$

可见只要能找到零均值随机变量 Y 使得 $\rho_{X,Y}\neq 0$ 就可以减小 EX 的估计方差。Y 和 X 的相关性越强，改善幅度越大。这种减小随机模拟误差方差的方法叫做**控制变量法**。

实际中 $\rho_{X,Y}$ 和 $\mathrm{Var}(X),\mathrm{Var}(Y)$ 可能是未知的，可以先模拟一个小的样本估计 $\rho_{X,Y}$ 和 $\mathrm{Var}(X),\mathrm{Var}(Y)$ 从而获得 b 的估计值。

控制变量法中要求控制变量 Y 与 X 相关且 $EY=0$。如果 $EY\neq 0$ 但 $EY=\mu_Y$ 已知，只要用 $Y-\mu_Y$ 代替 Y，这需要能预先知道 Y 的期望值的真值。另一种情况是 $EY=EX$ 未知，Y 与 X 相关，这时令 $Z=\alpha X+(1-\alpha)Y$, $\alpha\in[0,1]$, 则 $EZ=EX=\theta$，可以求 α 使得 $\mathrm{Var}(Z)$ 最小。容易知道当 $\alpha=\mathrm{Cov}(Y,Y-X)/\mathrm{Var}(Y-X)$ 时 $\mathrm{Var}(Z)$ 最小。

例 3.3.1. 设要估计 $I=\int_0^1 e^t\,dt$。当然，可以得到积分真值为 $e-1$，这里用来演示控制变量法的优势。

设 $U\sim\mathrm{U}(0,1)$，$X=e^U$, 则 $I=Ee^U=EX$, 可以用平均值法估计 I 为

$$\hat{I}_1=\frac{1}{N}\sum_{i=1}^{N}e^{U_i}. \tag{3.45}$$

其方差为

$$\mathrm{Var}(\hat{I}_1)=\frac{1}{N}\mathrm{Var}(e^U)=\frac{1}{N}(-\frac{1}{2}e^2+2e-\frac{3}{2})\approx\frac{0.2420}{N}. \tag{3.46}$$

令 $Y = U - \frac{1}{2}$, 则 $EY = 0$, X 与 Y 正相关，可以计算出 $\mathrm{Cov}(X, Y) \approx 0.14086$, $\mathrm{Var}(Y) = 1/12$ (更复杂的问题中可能需要从一个小的随机抽样中近似估计)，于是 $b = -\mathrm{Cov}(X, Y)/\mathrm{Var}(Y) = -1.690$, 对 $Z(b) = e^U - 1.690(U - \frac{1}{2})$ 有 $\mathrm{Var}(Z(b)) = [1 - \rho^2_{X,Y}]\mathrm{Var}(X) = (1 - 0.9919^2)\mathrm{Var}(X) = 0.016\mathrm{Var}(X) = 0.0039$, 用控制变量法估计 I 为

$$\hat{I}_2 = \frac{1}{N}\sum_{i=1}^{N}\left[e^{U_i} - 1.690(U_i - \frac{1}{2})\right].$$

$\hat{I}_1$ 的方差比控制变量法 $\hat{I}_2$ 的方差大 60 倍以上。 □

例 3.3.2. (系统可靠性估计) 考虑由 n 个部件组成的一个系统，用 S_i 表示第 i 个部件是否正常工作，1 表示正常工作，0 表示失效。设 $S_i \sim \mathrm{B}(1, p_i)$ 且各 S_i 相互独立。用 Y 表示系统是否工作正常，1 表示工作正常，0 表示系统失效。设 Y 为 $S_1, S_2, \ldots, S_n$ 的函数 $\phi(S_1, S_2, \ldots, S_n)$ 且 ϕ 关于每个 S_i 是单调不减，称 ϕ 为系统的结构函数。令 $R = P(Y = 1) = E\phi(S_1, S_2, \ldots, S_n)$，称 R 为系统可靠度。

例如，$\phi(s_1, s_2, \ldots, s_n) = \prod_{i=1}^{n} s_i$, 则当且仅当所有部件正常工作时系统才正常工作，这样的系统称为串联系统，这时系统可靠度为

$$R = P(S_1 = 1, S_2 = 1, \ldots, S_n = 1) = \prod_{i=1}^{n} P(S_i = 1) = p_1 p_2 \ldots p_n. \tag{3.47}$$

在系统比较简单的情况下 (如串联、并联)，可以给出用 $p_1, p_2, \ldots, p_n$ 表示 R 的表达式。但是更复杂的系统则很难写出 R 的表达式，这时可以用随机模拟方法估计 R。记 $\boldsymbol{S} = (S_1, S_2, \ldots, S_n)$, $X = \phi(\boldsymbol{S})$, 对 $\boldsymbol{S}$ 独立抽取 N 个点 $\boldsymbol{S}^{(j)} = (S_1^{(j)}, S_2^{(j)}, \ldots, S_n^{(j)}), j = 1, 2, \ldots, N$, R 可以用平均值法估计为

$$\hat{R}_1 = \frac{1}{N}\sum_{j=1}^{N}\phi(\boldsymbol{S}^{(j)}). \tag{3.48}$$

令 $Y = \sum_{i=1}^{n}(S_i - p_i)$, 则 $EY = 0$, Y 与 X 正相关。用一个小的抽样先近似估计 $\mathrm{Cov}(X, Y), \mathrm{Var}(Y)$ 得到 b 的近似值，可以得到方差缩减的估计量

$$\hat{R}_2 = \frac{1}{N}\sum_{j=1}^{N}\left[\phi(\boldsymbol{S}^{(j)}) + b\sum_{i=1}^{n}(S_i^{(j)} - p_i)\right]. \quad \square$$

3.3.2 对立变量法

控制变量法需要知道控制变量 Y 的期望值真值，并精确或近似知道 $\mathrm{Cov}(X, Y)$ 和 $\mathrm{Var}(Y)$。对立变量法的要求比较简单。

模拟中经常使用均匀分布随机数 U 变换产生的随机数 $X=g(U)$。下面的定理说明，如果变换 $g(\cdot)$ 是单调的，则随机变量 $Y=g(1-U)$ 就是与 X 负相关的。注意 $g(1-U)$ 与 $g(U)$ 同分布所以 $EY=EX$。

定理 3.3.1. 设 g 为单调函数，$U\sim \mathrm{U}(0,1)$, 则 $\mathrm{Cov}(g(U),g(1-U))\leq 0$。

证明. $\forall u_1,u_2\in[0,1]$，由 g 单调可知

$$(g(u_1)-g(u_2))\,(g(1-u_1)-g(1-u_2))\leq 0$$

设 U_2 服从 U(0,1) 且与 U 独立，令 $X_1=g(U),Y_1=g(1-U),\ X_2=g(U_2),Y_2=g(1-U_2)$, 则 X_1,Y_1,X_2,Y_2 的分布相同，且

$$\begin{aligned}
&E(X_1-X_2)(Y_1-Y_2)\\
=&\mathrm{Cov}(X_1-X_2,Y_1-Y_2)\\
=&\mathrm{Cov}(X_1,Y_1)+\mathrm{Cov}(X_2,Y_2)-\mathrm{Cov}(X_1,Y_2)-\mathrm{Cov}(X_2,Y_1)\\
=&2\mathrm{Cov}(X_1,Y_1)
\end{aligned}$$

注意 $(X_1-X_2)(Y_1-Y_2)\leq 0$ 所以 $\mathrm{Cov}(X_1,Y_1)\leq 0$，即 $\mathrm{Cov}(g(U),g(1-U))\leq 0$。证毕。 □

定理3.3.1可以推广到如下情形。

定理 3.3.2. 设 $h(x_1,x_2,\ldots,x_n)$ 是关于每个自变量单调的函数，$U_1,U_2,\ldots,U_n$ 相互独立，则 $\mathrm{Cov}(h(U_1,U_2,\ldots,U_n),\ h(1-U_1,1-U_2,\ldots,1-U_n))\leq 0$。

证明略去，参见 Ross(2013)[34] §9.9。

对均匀随机数 U 最常见的变换是逆变换 $X=F^{-1}(U)$。下面的定理给出了提高 $I=EX$ 估计精度的方法。

定理 3.3.3 (对立变量法). 设 $F(x)$ 为连续分布函数，$U\sim \mathrm{U}(0,1)$，$X=F^{-1}(U)$，$Y=F^{-1}(1-U)$，$Z=\frac{X+Y}{2}$，则 X 与 Y 同分布 $F(x)$ 且 $\mathrm{Cov}(X,Y)\leq 0$，

$$\mathrm{Var}(Z)\leq\frac{1}{2}\mathrm{Var}(X)$$

证明. 因为 U 和 $1-U$ 同分布所以 $X=F^{-1}(U)$ 和 $Y=F^{-1}(1-U)$ 同分布。由定理3.3.1, 令 $g(\cdot)=F^{-1}(\cdot)$ 可知 $\mathrm{Cov}(X,Y)=\mathrm{Cov}(g(U),g(1-U))\leq 0$，从而

$$\begin{aligned}
\mathrm{Var}(Z)=&\frac{\mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)}{4}\\
=&\frac{\mathrm{Var}(X)+\mathrm{Cov}(X,Y)}{2}\leq\frac{1}{2}\mathrm{Var}(X)
\end{aligned}$$

证毕。 □

根据定理3.3.3的结论，为了估计 $I=EX$，产生 $U_1,U_2,\ldots,U_N$ 后用

$$Z_i=\frac{1}{2}\left(F^{-1}(U_i)+F^{-1}(1-U_i)\right),\ i=1,2,\ldots,N$$

的样本平均值估计 I 可以提高精度, 在不增加抽样个数的条件下把估计的随机误差方差降低到原来的 $\frac{1}{2}$ 以下。这种提高随机模拟精度的方法叫做**对立变量法**。对立变量法不需要计算 X 和 Y 的方差及协方差的值，比控制变量法更简便易行。

例 3.3.3. 再次考虑例3.3.1的问题，估计 $I=\int_0^1 e^t\,dt$。下面用对立变量法改善原始的平均值法 $\hat{I}_1$ 的估计方差。

设 $U\sim \mathrm{U}(0,1)$，$X=e^U$, 令 $Y=e^{1-U}$, 用对立变量法估计 I 为

$$\hat{I}_3=\frac{1}{N}\sum_{i=1}^{N}\frac{e^{U_i}+e^{1-U_i}}{2}, \tag{3.49}$$

方差为

$$\mathrm{Var}(\hat{I}_3)=\frac{1}{N}\frac{\mathrm{Var}(e^U)+\mathrm{Cov}(e^U,e^{1-U})}{2}=\frac{1}{N}(-\frac{3}{4}e^2+\frac{5}{2}e-\frac{5}{4})\approx\frac{0.003913}{N}, \tag{3.50}$$

$\hat{I}_1$ 的方差比对立变量法估计 $\hat{I}_3$ 的方差大至少 60 倍，而 $\hat{I}_3$ 的方差和例3.3.1中控制变量法估计量 $\hat{I}_2$ 的方差相近。 □

对立变量法和控制变量法是类似做法，一般不能结合使用。

例 3.3.4. 再次考虑例3.3.2的可靠度估计问题。用对立变量法改善估计方差。设 $\{U_k\}$ 为标准均匀分布随机数列，取

$$S_i^{(j)}=\begin{cases}1 & \text{当}U_{n(j-1)+i}\le p_i,\\ 0 & \text{其它}\end{cases} \tag{3.51}$$

则 $S_i^{(j)}$ 是 $U_{n(j-1)+i}$ 的单调非增函数。R 用平均值法估计为

$$\hat{R}_1=\frac{1}{N}\sum_{j=1}^{N}\phi(S_1^{(j)},S_2^{(j)},\ldots,S_n^{(j)}). \tag{3.52}$$

利用对立变量法，令 $h(U_1,U_2,\ldots,U_n)=\phi(S_1,S_2,\ldots,S_n)$, 则 h 关于每个自变量是单调非增函数，于是

$$\mathrm{Cov}(h(U_1,U_2,\ldots,U_n),\ h(1-U_1,1-U_2,\ldots,1-U_n))\le 0, \tag{3.53}$$

估计系统可靠度 R 为

$$\hat{R}_3 = \frac{1}{N}\sum_{j=1}^{N}\frac{h(U_1^{(j)},U_2^{(j)},\ldots,U_n^{(j)}) + h(1-U_1^{(j)},1-U_2^{(j)},\ldots,1-U_n^{(j)})}{2} \tag{3.54}$$

就能比 $\hat{R}_1$ 的误差方差至少降低 $\frac{1}{2}$。 □

注意定理3.3.2中的 $U_1,U_2,\ldots,U_n$ 仅要求独立，并未指定分布。事实上，此结果还可以推广。如果 $(X_i,Y_i), i=1,2,\ldots,n$ 独立，且 X_i 和 Y_i 负相关，$h(x_1,x_2,\ldots,x_n)$ 关于每个自变量单调不减，则 $\mathrm{Cov}(h(X_1,X_2,\ldots,X_n),\ h(Y_1,Y_2,\ldots,Y_n)) \le 0$，可以类似地用控制变量法或对立变量法的做法构造方差缩减估计量。例如，设 $X_i \sim \mathrm{N}(\mu_i,\sigma_i^2)$ 相互独立，要估计 $\theta = Eh(X_1,X_2,\ldots,X_n)$, 其中 h 关于每个自变量单调不减，则 $h(2\mu_1-X_1,2\mu_2-X_2,\ldots,2\mu_n-X_n)$ 与 $h(X_1,X_2,\ldots,X_n)$ 同分布，对

$$Z = \frac{h(X_1,X_2,\ldots,X_n) + h(2\mu_1-X_1,2\mu_2-X_2,\ldots,2\mu_n-X_n)}{2}$$

抽样用平均值估计 θ 可以比仅对 $h(X_1,X_2,\ldots,X_n)$ 抽样得到的估计量的方差缩减一半以上。

3.3.3 条件期望法

进行统计估计时，如果有额外的相关信息，利用这样的信息可以提高估计精度。比如，对随机变量 Y，如果 Y 服从某种模型，在估计 $I=EY$ 时应当尽量利用模型信息。

设变量 X 与 Y 不独立，根据 Rao-Blackwell 不等式:

$$\mathrm{Var}\{\mathrm{E}(Y|X)\} \le \mathrm{Var}(Y)$$

又

$$\mathrm{E}\{\mathrm{E}(Y|X)\} = \mathrm{E}Y = I$$

所以，对 $Z=\mathrm{E}(Y|X)$ 抽样, 用 Z 的样本平均值来估计 $I=\mathrm{E}Y$ 比直接用 Y 的样本平均值的精度更高。这种改善随机模拟估计精度的方法叫做**条件期望法**，或 Rao-Blackwell 方法。

例 3.3.5. 设 $X \sim p_1(x)$，$\varepsilon \sim \mathrm{N}(0,\sigma^2)$ 且与 X 独立，

$$Y = \psi(X) + \varepsilon,$$

要估计 $I=\mathrm{E}Y$, 可以用条件分布抽样法对二元随机向量 $\boldsymbol{Z}=(X,Y)$ 抽样产生 Y 的样本。从 $p(\cdot)$ 抽样得 $X_1,X_2,\ldots,X_N$, 独立地从 $\mathrm{N}(0,\sigma^2)$ 抽样得 $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N$，令

$$Y_i = \psi(X_i) + \varepsilon_i,\ i=1,2,\ldots,N$$

然后用 $Y_1, Y_2, \ldots, Y_N$ 的样本平均值估计 EY:

$$\hat{I}_1 = \frac{1}{N}\sum_{i=1}^{N} Y_i,$$

估计方差为

$$\mathrm{Var}(\hat{I}_1) = \frac{\mathrm{Var}(Y_1)}{N} = \frac{\mathrm{Var}(\psi(X_1))}{N} + \frac{\sigma^2}{N}.$$

另一方面，注意 $E(Y|X) = \psi(X)$，也可以只对 X 抽样然后用条件期望法估计 EY:

$$\hat{I}_2 = \frac{1}{N}\sum_{i} \psi(X_i),$$

估计方差为

$$\mathrm{Var}(\hat{I}_2) = \frac{\mathrm{Var}(\psi(X_1))}{N} < \mathrm{Var}(\hat{I}_1).$$

这个例子演示了条件期望法可以缩减误差方差的原因：对 Y 抽样分成两步进行：第一步对 X 抽样，第二步利用 $Y|X$ 的条件分布对 Y 抽样。所以使用 Y 的样本估计 EY 包含了第二步对 Y 抽样的随机误差，而使用 X 的函数 $E(Y|X) = \psi(X)$ 的抽样来估计 EY 则避免了第二步对 Y 抽样引起的随机误差。 □

例 3.3.6. 继续考虑例3.1.1中用随机模拟方法估计 π 的问题。设 (X, Y) 服从正方形 $D = [-1, 1]^2$ 上的均匀分布，令 $\eta = 1$ 表示 (X, Y) 落入单位圆 $C = \{(x, y) : x^2 + y^2 \le 1\}$, $\eta = 0$ 表示未落入单位圆, 则 $\eta \sim \mathrm{B}(1, \pi/4)$。设 $(X_i, Y_i), i = 1, 2, \ldots, N$ 是 (X, Y) 的独立重复抽样，η_i 表示 $X_i^2 + Y_i^2 \le 1$ 是否发生，则 π 的估计为

$$\hat{\pi}_1 = \frac{4}{N}\sum_{i=1}^{N} \eta_i,$$

方差为 $\mathrm{Var}(\hat{\pi}_1) = \pi(4 - \pi)/N \approx \frac{2.6968}{N}$。

用 $\zeta = E(\eta|X)$ 的样本代替 η 来估计 $E\eta = \pi/4$，则由

$$\begin{aligned} E(\eta|X = x) =& P(X^2 + Y^2 \le 1 \,|\, X = x) = P(Y^2 \le 1 - x^2) \\ =& P(-\sqrt{1 - x^2} \le Y \le \sqrt{1 - x^2}) \\ =& \sqrt{1 - x^2} \quad (\text{注意} Y \sim \mathrm{U}(-1, 1)) \end{aligned}$$

可知 $\zeta = \sqrt{1 - X^2}$。其方差为

$$\mathrm{Var}(\zeta) = E(1 - X^2) - (E\zeta)^2 = \frac{2}{3} - \frac{\pi^2}{16},$$

估计 π 为

$$\hat{\pi}_2 = \frac{4}{N}\sum_{i=1}^{N}\sqrt{1-X_i^2},$$

方差为 $\mathrm{Var}(\hat{\pi}_2) \approx 0.7971/N$，$\mathrm{Var}(\hat{\pi}_1)/\mathrm{Var}(\hat{\pi}_2) \approx 3.4$。

另外，ζ 仅依赖于 X^2，容易发现若 $U \sim \mathrm{U}(0,1)$ 则 X^2 和 U^2 同分布，所以可取 $\xi = \sqrt{1-U^2}$, 其中 $U \sim \mathrm{U}(0,1)$。这时，函数 $h(u) = \sqrt{1-u^2}$ 是 $u \in (0,1)$ 的单调函数，可以利用对立变量法，构造 π 的估计量为

$$\hat{\pi}_3 = \frac{4}{N}\sum_{i=1}^{N}\frac{\sqrt{1-U_i^2}+\sqrt{1-(1-U_i)^2}}{2},$$

其中 $U_1, U_2, \ldots, U_N$ 为 $U(0,1)$ 随机数，则 $\mathrm{Var}(\hat{\pi}_3) \approx 0.11/N$, $\mathrm{Var}(\hat{\pi}_1)/\mathrm{Var}(\hat{\pi}_3) \approx 25$。

也可以用对立变量法来改进 $\hat{\pi}_2$。令 $W = U^2 - \frac{1}{3}$, 则 $EW = 0$ 且 W 与 ζ 负相关。可以先进行一个小规模的模拟估计 $\mathrm{Cov}(\zeta, W)$ 和 $\mathrm{Var}(W)$ 得到 $b = -\mathrm{Cov}(\zeta, W)/\mathrm{Var}(W)$ 的近似值，用 $\zeta(b) = \zeta + bW$ 代替 ζ 进行抽样，可以减小 $\hat{\pi}_2$ 的方差。 □

例 3.3.7. 设随机变量 $\boldsymbol{Y} \sim f(\boldsymbol{y})$, 为了估计 $\theta = Eh(\boldsymbol{Y})$, 经常使用标准化重要抽样法

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{N} W_i h(\boldsymbol{X}_i)}{\sum_{i=1}^{N} W_i}, \tag{3.55}$$

其中 $\boldsymbol{X}_i$ 为试抽样密度 $g(\boldsymbol{x})$ 的样本，权重 $W_i = f(\boldsymbol{X}_i)/g(\boldsymbol{X}_i)$。如果 $\boldsymbol{x} = (\boldsymbol{u}, \boldsymbol{v})$, $h(\boldsymbol{x}) = h_1(\boldsymbol{u})$, 则只要对 $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{V})$ 的分量 $\boldsymbol{U}$ 抽样得到 $\boldsymbol{U}_i, i = 1, \ldots, N$ 及新的权重 $\tilde{W}_i = f_{\boldsymbol{U}}(\boldsymbol{U}_i)/g_{\boldsymbol{U}}(\boldsymbol{U}_i)$, 其中 $f_{\boldsymbol{U}}(\boldsymbol{u}) = \int f(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{v}$, $g_{\boldsymbol{U}}(\boldsymbol{u}) = \int g(\boldsymbol{u}, \boldsymbol{v})\, d\boldsymbol{v}$, 则可估计 θ 为

$$\hat{\theta}_2 = \frac{\sum_{i=1}^{N} \tilde{W}_i h_1(\boldsymbol{U}_i)}{\sum_{i=1}^{N} \tilde{W}_i}, \tag{3.56}$$

这时

$$\mathrm{Var}(\tilde{W}_i) \le \mathrm{Var}(W_i),\ i = 1, 2, \ldots, N. \tag{3.57}$$

事实上，

$$\begin{aligned}\frac{f_{\boldsymbol{U}}(\boldsymbol{u})}{g_{\boldsymbol{U}}(\boldsymbol{u})} &= \int \frac{f(\boldsymbol{u}, \boldsymbol{v})}{g_{\boldsymbol{U}}(\boldsymbol{u})}\, dv = \int \frac{f(\boldsymbol{u}, \boldsymbol{v})}{g_{\boldsymbol{U}}(\boldsymbol{u}) g_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{u})} g_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{u})\, d\boldsymbol{v} \\ &= \int \frac{f(\boldsymbol{u}, \boldsymbol{v})}{g(\boldsymbol{u}, \boldsymbol{v})} g_{\boldsymbol{V}|\boldsymbol{U}}(\boldsymbol{v}|\boldsymbol{u})\, d\boldsymbol{v} = E_g\left\{ \frac{f(\boldsymbol{U}, \boldsymbol{V})}{g(\boldsymbol{U}, \boldsymbol{V})} \middle| \boldsymbol{U} = \boldsymbol{u} \right\}\end{aligned} \tag{3.58}$$

于是

$$\mathrm{Var}_g\left\{ \frac{f(\boldsymbol{U}, \boldsymbol{V})}{g(\boldsymbol{U}, \boldsymbol{V})} \right\} \ge \mathrm{Var}_g\left\{ E_g\left[\frac{f(\boldsymbol{U}, \boldsymbol{V})}{g(\boldsymbol{U}, \boldsymbol{V})} \middle| \boldsymbol{U} \right] \right\} = \mathrm{Var}_g\left\{ \frac{f_{\boldsymbol{U}}(\boldsymbol{U})}{g_{\boldsymbol{U}}(\boldsymbol{U})} \right\}. \tag{3.59}$$

□

3.3.4 随机数复用

在统计研究中，经常需要比较若干种统计方法的性能，如偏差、方差、覆盖率等。除了努力获取理论结果以外，可以用随机模拟方法进行比较：重复生成 N 组随机样本，对每个样本同时用不同统计方法计算结果，最后从 N 组结果比较不同方法的性能。这样比较时，并没有对每种方法单独生成 N 组样本，而是每个样本同时应用所有要比较的方法。这样不仅减少了计算量，而且在比较时具有更高的精度。

例 3.3.8. 对正态分布总体 $X \sim N(\mu, \sigma^2)$, 如果有样本 $X_1, X_2, \ldots, X_n$, 估计 σ^2 有两种不同的公式:

$$\hat{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{3.60}$$

希望比较两个估计量的偏差和均方误差:

$$b_1 = E\hat{\sigma}_1^2 - \sigma^2, \quad b_2 = E\hat{\sigma}_2^2 - \sigma^2, \tag{3.61}$$

$$s_1 = E(\hat{\sigma}_1^2 - \sigma^2)^2, \quad s_2 = E(\hat{\sigma}_2^2 - \sigma^2)^2. \tag{3.62}$$

当然，这个问题很简单，可以得到偏差和均方误差的理论值：

$$b_1 = 0, \quad b_2 = -\frac{1}{n}\sigma^2, \tag{3.63}$$

$$s_1 = \frac{2\sigma^4}{n-1}, \quad s_2 = \frac{(2n-1)\sigma^4}{n^2}, \quad s_1 - s_2 = \frac{(3n-1)\sigma^4}{n^2(n-1)}. \tag{3.64}$$

我们用随机模拟来作比较。重复地生成 N 组样本 $(X_1^{(j)}, X_2^{(j)}, \ldots, X_n^{(j)})$, $j = 1, 2, \ldots, N$。对每组样本分别计算 $\hat{\sigma}_{1j}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^{(j)} - \bar{X}^{(j)})^2$ 和 $\hat{\sigma}_{2j}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^{(j)} - \bar{X}^{(j)})^2$，得到偏差和均方误差的估计

$$\hat{b}_1 = \frac{1}{N}\sum_{j=1}^{N}\hat{\sigma}_{1j}^2 - \sigma^2, \quad \hat{b}_2 = \frac{1}{N}\sum_{j=1}^{N}\hat{\sigma}_{2j}^2 - \sigma^2, \tag{3.65}$$

$$\hat{s}_1 = \frac{1}{N}\sum_{j=1}^{N}(\hat{\sigma}_{1j}^2 - \sigma^2)^2, \quad \hat{s}_2 = \frac{1}{N}\sum_{j=1}^{N}(\hat{\sigma}_{2j}^2 - \sigma^2)^2. \tag{3.66}$$

容易看出，两种方法使用相同的模拟样本得到的偏差、均方误差的估计精度与每种方法单独生成模拟样本得到的估计精度相同。

但是，如果要估计 $\Delta s = s_1 - s_2$，利用相同的样本的估计精度更好。一般地，

$$\mathrm{Var}(\hat{s}_1 - \hat{s}_2) = \mathrm{Var}(\hat{s}_1) + \mathrm{Var}(\hat{s}_2) - 2\mathrm{Cov}(\hat{s}_1, \hat{s}_2). \tag{3.67}$$

如果每种方法使用不同的样本，则(3.67)变成

$$\mathrm{Var}(\hat{s}_1-\hat{s}_2)=\frac{1}{N}\left[\mathrm{Var}((\hat{\sigma}_1^2-\sigma^2)^2)+\mathrm{Var}((\hat{\sigma}_2^2-\sigma^2)^2)\right]. \tag{3.68}$$

如果两种方法利用相同的样本计算，则(3.67) 变成

$$\mathrm{Var}(\hat{s}_1-\hat{s}_2)=\frac{1}{N}\left[\mathrm{Var}((\hat{\sigma}_1^2-\sigma^2)^2)+\mathrm{Var}((\hat{\sigma}_2^2-\sigma^2)^2)-2\mathrm{Cov}((\hat{\sigma}_1^2-\sigma^2)^2,(\hat{\sigma}_2^2-\sigma^2)^2)\right]. \tag{3.69}$$

而 $(\hat{\sigma}_1^2-\sigma^2)^2$ 与 $(\hat{\sigma}_2^2-\sigma^2)^2$ 明显具有强正相关 (见习题10)，所以两种方法针对相同样本计算时 $\hat{s}_1-\hat{s}_2$ 的方差比两种方法使用单独样本时的方差要小得多。这说明重复利用相同的随机数或样本往往可以提高比较的精度。 □

3.4 随机服务系统模拟 *

我们在 §3.1中讲到，较为复杂的随机模型往往难以进行彻底的理论分析，这时常常使用随机模拟方法产生模型的大量数据，从产生的数据对模型进行统计推断。随机服务系统就是这样的一种模型，经常需要利用随机模拟方法进行研究。

随机服务系统在我们日常生活、工业生产、科学技术、军事领域中是经常遇到的随机模型，比如，研究银行、理发店、商店、海关通道、高速路收费口等服务人员个数的设置和排队规则，研究计算机网络网关、移动网络的调度规则，等等。

在概率统计理论中排队论用来研究随机服务系统的数学模型，可以用来设计适当的服务机构数目和排队规则。如下面的 M/M/1 排队系统。

例 3.4.1. 设某银行仅有一个柜员，并简单假设银行不休息。顾客到来间隔的时间服从独立的指数分布 $\mathrm{Exp}(\lambda)$ ($1/\lambda$ 为间隔时间的期望值), 如果柜员正在为先前的顾客服务, 新到顾客就排队等待，柜员为顾客服务的时间服从均值为 $1/\mu$ 的指数分布，设 $u=\lambda/\mu<1$。设 X_t 表示 t 时刻在银行内的顾客数（包括正在服务的和正在排队的），则 X_t 是一个连续时马氏链。

这是一个生灭过程马氏链，有理论结果当系统处于稳定状态时

$$P(X_t=i)=u^i(1-u),\ i=0,1,2,\ldots.$$

设随机变量 N 服从 X_t 的平稳分布。于是银行中平均顾客数为

$$EN=\frac{u}{1-u},$$

平均队列长度 EQ 等于 EN 减去平均正在服务人数，正在服务人数 Y_t 为

$$Y_t = \begin{cases} 1, & \text{当} X_t > 0, \\ 0, & \text{当} X_t = 0 \end{cases}$$

所以 $EY_t = P(Y_t = 1) = 1 - P(N = 0) = u$, 于是平均队列长度为

$$EQ = EN - EY_t = \frac{u}{1-u} - u = \frac{u^2}{1-u},$$

设顾客平均滞留时间为 ER，由关系式

$$EN = \lambda \cdot ER$$

可知平均滞留时间为

$$ER = \frac{u}{\lambda(1-u)} = \frac{1}{\mu - \lambda},$$

进一步分析还可以知道顾客滞留时间 R 服从均值为 $1/(\mu - \lambda)$ 的指数分布。

从上面的例子可以发现，一个随机服务系统的模型应该包括如下三个要素：

- 输入过程: 比如，银行的顾客到来的规律。
- 排队规则: 比如，银行有多个柜员时，顾客是选最短一队，还是随机选一队，还是统一排队等候叫号，顾客等得时间过长后是否会以一定概率放弃排队。
- 服务机构：有多少个柜员，服务时间的分布等。

虽然某些随机服务系统可以进行严格理论分析得到各种问题的理论解，但是，随机服务系统中存在大量随机因素，使得理论分析变得很困难以至于不可能。例如，即使是上面的银行服务问题，可能的变化因素就包括：顾客到来用齐次泊松过程还是非齐次泊松过程，柜员有多少个，是否不同时间段柜员个数有变化，柜员服务时间服从什么样的分布，顾客排队按照什么规则，是否 VIP 顾客提前服务，顾客等候过长时会不会放弃排队，等等。包含了这么多复杂因素的随机服务系统的理论分析会变得异常复杂，完全靠理论分析无法解决问题，这时，可以用随机模拟方法给出答案。

在模拟随机服务系统时，可以按时间顺序记录发生的事件，如顾客到来、顾客接受服务、顾客结束服务等，这样的系统的模拟也叫做**离散事件模拟**。

离散事件模拟算法可以分为三类：

- 活动模拟，把连续时间离散化为很小的单位，比如平均几秒发生一个新事件时可以把时间精确到毫秒，然后时钟每隔一个小时间单位前进一步并查看所有活动并据此更新系统状态。缺点是速度太慢。

- 事件模拟。仅在事件发生时更新时钟和待发生的事件集合。这样的方法不受编程语言功能的限制，运行速度很快，也比较灵活，可以实现复杂逻辑，但是需要自己管理的数据结构和逻辑结构比较复杂，算法编制相对较难。

- 过程模拟。把将要到来的各种事件看成是不同的过程，让不同的过程并行地发生, 过程之间可以交换消息，并在特殊的软件包或编程语言支持下自动更新系统状态。在计算机实现中需要借助于线程或与线程相似的程序功能。这类软件包有 C++ 语言软件包 C++SIM 和 Python 语言软件包 SimPy。优点是系统逻辑的编码很直观，程序模块化，需要用户自己管理的数据结构和逻辑结构少。过程模拟是现在更受欢迎的离散事件模拟方式。

事件模拟算法必须考虑的变量包括：

- 当前时刻 t;

- 随时间而变化的计数变量，如 t 时刻时到来顾客人数、已离开人数；

- 系统状态，比如是否有顾客正在接受服务、排队人数、队列中顾客序号。

这些变量仅在有事件发生时（如顾客到来、顾客离开）才需要记录并更新。为了持续模拟后续事件，需要维护一个将要发生的事件的列表（下一个到来时刻、下一个离开时刻），列表仅需要在事件发生时进行更新。其它变量可以从这三种变量中推算出来，比如，顾客 i 在时间 t_1 时到达并排队，在时间 t_2 时开始接受服务，在时间 t_4 时结束服务离开，则顾客 i 的滞留时间为 $t_4 - t_1$。

例 3.4.2. 用事件模拟的方法来模拟例3.4.1。目的是估计平均滞留时间 ER。想法是，模拟生成各种事件发生的时间，模拟很长时间，丢弃开始的一段时间 T_0 后，用 T_0 后到达的顾客的总滞留时间除以 T_0 后到达的顾客人数来估计平均滞留时间。设 T_0 时间后每位顾客滞留时间的模拟值为 $R_i, i = 1, 2, \ldots, m$，可以用 $\{R_i\}$ 作为随机变量 R 的样本来检验 R 的分布是否指数分布。

用事件模拟方法进行离散事件模拟的算法关键在于计算系统状态改变的时间，即各个事件的发生时间，这个例子中就是顾客到来、顾客开始接受服务、顾客离开这样三种事件, 由此还可以得到每个顾客排队的时间和服务的时间。

在没有明确算法构思时，可以从时间 0 开始在纸上按照模型规定人为地生成一些事件并人为地找到下一事件。这样可以找到要更新的数据结构和更新的程序逻辑。

下面的算法保持了一个将要发生的事件的集合，在每次事件发生时更新时钟，更新时钟时从事件集合中找到最早发生的事件进行处理，并生成下一事件到事件集合中，如此重复直到需要模拟的时间长度。

{初始化当前时钟 $t \leftarrow 0$, 柜员忙标志 $B \leftarrow 0$, 当前排队人数 $L \leftarrow 0$,
最新到来顾客序号 $i \leftarrow 0$, 正在服务顾客序号 $j \leftarrow 0$, 已服务顾客数 $n \leftarrow 0$ }
从 Exp(λ) 抽取 X, 设置下一顾客来到时间 $A \leftarrow X$
repeat {
  if($B = 0$ **or** ($B = 1$ **and** $A < E$)) { # E 是正在服务的顾客结束时刻
    $t \leftarrow A$
  } **else** {
    $t \leftarrow E$
  }
  if ($t > T_1$) **break** # T_1 是预先确定的模拟时长

  if($t == A$) { # 待处理到达事件
    $L \leftarrow L + 1$
    $i \leftarrow i + 1$, 记录第 i 位顾客到来时间 $a_i \leftarrow t$
    从 Exp(λ) 抽取 X, $A \leftarrow t + X$
    if($B = 0$) { # 不用排队，直接服务
      $B \leftarrow 1$, $L \leftarrow L - 1$
      $j \leftarrow j + 1$, 置第 j 位顾客开始服务时间 $s_j \leftarrow t$
      从 Exp(μ) 抽取 Y, 置 $E \leftarrow t + Y$
    }
  } **else** { # 待处理结束服务事件
    $B \leftarrow 0$
    $n \leftarrow n + 1$, 记录第 n 个顾客结束服务时间 $e_n \leftarrow t$
    if($L > 0$) { # 排队顾客开始服务
      $L \leftarrow L - 1$
      $B \leftarrow 1$
      $j \leftarrow j + 1$, $s_j \leftarrow t$

从 Exp(μ) 抽取 Y, 置 $E \leftarrow t + Y$

}

}

}

{ 令 $I = \{i:\ T_0 \le s_i \le T_1\}$, 求 $\{e_i - a_i,\ i \in I\}$ 的平均值作为 ER 估计 }

从这个算法可以看出，事件模拟方法需要用户自己管理待处理事件集合与时钟, 算法设计难度较大。

用随机服务系统进行建模和模拟研究的一般步骤如下：

- 提出问题。比如，银行中顾客平均滞留时间与相关参数的关系。
- 建立模型。比如，设顾客到来服从泊松过程，设每个顾客服务时间服从独立的指数分布。
- 数据收集与处理。比如，估计银行顾客到来速率，每个顾客平均服务时间，等等。
- 建立模拟程序。一般使用专用的模拟软件包或专用模拟语言编程。
- 模拟模型的正确性确认。
- 模拟试验和模拟结果分析。

离散事件模拟问题一般都比较复杂，即使借助于专用模拟软件或软件包，也很难确保实现的模拟算法与问题的实际设定是一致的。为此，需要遵循一些提高算法可信度的一般规则。算法程序一定要仔细检查，避免出现参数错误、逻辑错误。在开始阶段，可以利用较详尽的输出跟踪模拟运行一段时间，人工检查系统运行符合原始设定。尽可能利用模块化程序设计，比如，在 M/M/1 问题模拟中，顾客到来可能遵循不同的规律，比如时齐泊松过程、非时齐泊松过程，把产生顾客到来时刻的程序片段模块化并单独检查验证，就可以避免在这部分出错。问题实际设定可能比较复杂，在程序模块化以后，如果有一种较简单的设定可以得到理论结果，就可以用理论结果验证算法输出，保证程序框架正确后，再利用模块化设计修改各个模块适应实际复杂设定。

3.5 统计研究与随机模拟

现代统计学期刊发表的论文中一半以上包括随机模拟结果（也叫数值结果）。随机模拟可以辅助说明新模型或新方法的有效性，尤其是在很难获得关于有效性的理论结果的情况下让我们也能说明自己的新方法的优点。

比如，为了掌握统计量的抽样分布经常需要依靠随机模拟。在独立正态样本情况下，我们已经有样本均值和样本方差的分布，但是对其它分布 $F(x)$ 的样本，我们一般只能得到统计量 $\hat{\theta}$ 的大样本性质，在中小样本情况下很难得到 $\hat{\theta}$ 的抽样分布理论结果，随机模拟可以解决这样的问题。在抽取 $F(x)$ 的很多批样本后，得到很多个统计量 $\hat{\theta}$ 样本值，从这些样本值可以估计 $\hat{\theta}$ 的抽样分布，并研究样本量多大时 $\hat{\theta}$ 的抽样分布与大样本极限分布相符，计算估计的均方误差，等等。

下面举一个置信区间比较的例子说明统计研究中随机模拟的做法。实际应用中经常需要计算某个百分比的置信区间，比如北京市成年人口中希望房价降低的人的百分比的置信区间。

设随机抽取了样本量为 n 样本，其中成功个数为 X 个，则 $X \sim \mathrm{B}(n,p)$, p 为成功概率，要计算 p 的置信度为 $1-\alpha$ 的置信区间。记样本中成功百分比为 $\hat{p} = X/n$。假设样本量 $n \geq 30$。

当 n 较大时由中心极限定理可知

$$W = \frac{\hat{p} - p}{\sqrt{p(1-p)}/\sqrt{n}} \tag{3.70}$$

近似服从标准正态分布。因为 $\hat{p} \to p$ a.s. $(n \to \infty)$, 所以把(3.70)式分母中的 p 换成 $\hat{p}$，仍有

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}$$

近似服从标准正态分布。于是可以构造 p 的置信度 $1-\alpha$ 的置信区间为

$$\hat{p} \pm \frac{\lambda}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})} \tag{3.71}$$

其中 $\lambda = z_{1-\frac{\alpha}{2}}$ 是标准正态分布的 $1-\frac{\alpha}{2}$ 分位数。当 n 很大时这个置信区间应该是比较精确的，即 p 落入置信区间的实际概率应该很接近于标称的置信度 $1-\alpha$，但是在 n 不是太大的时候，p 落入置信区间的实际概率就可能与标称的置信度 $1-\alpha$ 有一定的差距，我们要用随机模拟考察 $1-\alpha, n, p$ 的不同取值下置信区间(3.71)的覆盖率。称置信区间实际包含真实参数的概率为**覆盖率**。

参数 p 的另一个置信区间, 称为 Wilson 置信区间，是直接求解关于 p 求解不等式

$$\left| \frac{\hat{p} - p}{\sqrt{p(1-p)}/\sqrt{n}} \right| \leq \lambda. \tag{3.72}$$

记

$$\tilde{p} = \frac{\hat{p} + \frac{\lambda^2}{2n}}{1 + \frac{\lambda^2}{n}}, \quad \delta = \frac{\lambda}{\sqrt{n}} \frac{\sqrt{\hat{p}(1-\hat{p}) + \frac{\lambda^2}{4n}}}{1 + \frac{\lambda^2}{n}},$$

则求解(3.72)得到的 p 的置信区间为

$$\tilde{p} \pm \delta \tag{3.73}$$

我们用随机模拟来考察(3.71)和(3.73)在不同 n, p 下的覆盖率并比较这两种置信区间。

好的置信区间应该满足如下两条要求：

(1) 覆盖率大于等于置信度，两者越接近越好；

(2) 置信区间越短越好。

下面设计这个模拟试验。设总共重复试验 M 次。M 越大，对覆盖率和置信区间长度期望估计的随机误差越小。比如我们希望模拟计算的覆盖率误差控制在 0.1% 以下 (在 95% 置信度下)。对每组样本，真值是否落入计算得到的置信区间是一个成败型试验，设真实覆盖率为 r，M 次重复试验中置信区间包含真实参数的次数为 V，则 $V \sim \mathrm{B}(M, r)$，覆盖率的估计值为 V/M，此估计的标准误差为 $\sqrt{r(1-r)}/\sqrt{M}$。如果覆盖率 r 近似等于置信区间的置信度 $1-\alpha$，则若 $1-\alpha \geq 0.8$，近似地有 $r(1-r) \leq 0.8 \times 0.2 = 0.16$, M 次试验的覆盖率估计的标准误差为 $\sqrt{0.16}/\sqrt{M} = 0.4/\sqrt{M}$, 以 2 倍标准误差作为覆盖率估计的误差大小界限（根据中心极限定理可知有 95% 的概率使得 $|\frac{V}{M} - r|$ 小于 2 倍标准误差），令 $2 \times 0.4/\sqrt{M} \leq 0.001$ 则有 $M = 640000$，这个试验重复量很大，如果程序耗时过长我们只好降低对随机误差幅度的要求。对于更复杂的问题，如果难以得到所需重复次数 M 的理论值，可以逐步增加 M，直到结果在需要的精度上不再变化为止。

比较(3.71)和(3.73)要考虑多种 $(n, p, 1-\alpha)$ 组合的影响。取 $1-\alpha = 0.99, 0.95, 0.90, 0.80$ 几种。(n, p) 的值要考虑到各种不同情况，初步选取如下的一些 (n, p) 组合:

$$\begin{aligned}
&n = 30, p = 0.1, 0.3, 0.5, 0.7, 0.9;\\
&n = 120, p = 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95\\
&n = 480, p = 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9, 0.95, 0.99
\end{aligned}$$

这样，我们设计了 $(n, p, 1-\alpha)$ 的 $4 \times (5+7+9) = 84$ 种不同情况，每种可以得到四个数值：置信区间(3.71)是否包含参数真值 (1 为包含，0 为不包含), 置信区间(3.73)是否包含参数真值, 置信区间(3.71)的长度, 置信区间(3.73)的长度。对每种给定 $(n, p, 1-\alpha)$ 的情况，我们重复试验 M 次，分别得到置信区间(3.71)的覆盖率 $\hat{r}_1$, 置信区间(3.73)的覆盖率 $\hat{r}_2$, 置信区间(3.71)的平均长度 $\bar{l}_1$ 和长度标准差 s_1, 置信区间(3.73)的平均长度 $\bar{l}_2$ 和长度标准差 s_2。

实际编程时，我们先编写 84 种不同情况中的一种情况去试验，首先试验较少的重复数，比如 $M = 100$。当程序检查无误后，再逐步增大重复次数看计算时间和存储能力是否可以

接受。对此例发现 $M = 640000$ 可行。最后，用循环处理所有 84 种情况, 每种情况重复试验 $M = 640000$ 次。最后的结果汇总成表3.1(限于篇幅，结果有删减)。表中每一行是一种 $(1-\alpha, n, p)$ 的组合。注意, $\bar{l}_1$ 和 $\bar{l}_2$ 的精度可以用其标准误差 $\mathrm{SE}_1 = s_1/\sqrt{M}$ 和 $\mathrm{SE}_2 = s_2/\sqrt{M}$ 来估计。

从表3.1的结果看出，即使样本量已经很大 $(n = 480)$，公式(3.71)的覆盖率仍然会低于置信度 $1-\alpha$，特别是当 p 靠近 0 或 1 的情况下公式(3.71)给出的置信区间更差。另一方面，公式(3.73)的覆盖率除少数例外总是略高于标称的置信度而且与标称置信度差距一般不大，两个公式得到的置信区间的平均长度相近。但是，$1-\alpha = 0.8, n = 480, p = 0.01$ 的 Wilson 置信区间覆盖率只有 74.7%，说明 Wilson 置信区间仍有改进必要。

3.6 Bootstrap 方法 *

3.6.1 标准误差

在统计建模中，伴随着参数的估计值，应该同时给出估计的“标准误差”。设总体 $X \sim F(x,\theta)$, $\theta \in \Theta$，$\hat{\phi}$ 是总体的一个参数 ϕ 的估计量，称 $\mathrm{SE} = \sqrt{\mathrm{Var}(\hat{\phi})}$ 为 $\hat{\phi}$ 的**标准误差**。实际工作中 SE 一般是未知的，SE 的估计也称为 $\hat{\phi}$ 的标准误差。对有偏估计，除了标准误差外我们还希望能够估计偏差。进一步地，我们还可能希望得到统计量 $\hat{\phi}$ 的分布，称为抽样分布。

例 3.6.1. 设 $X_i, i = 1, \ldots, n$ 是总体 $X \sim F(x)$ 的样本, 样本平均值 $\hat{\phi} = \bar{X} = \frac{1}{n}\sum_i X_i$ 为 $\phi = EX$ 的点估计，$\mathrm{SE}(\bar{X}) = \sqrt{\mathrm{Var}(X)/n}$，可以用 $S/\sqrt{n}$ 估计 $\mathrm{SE}(S^2$ 为样本方差)。根据中心极限定理和强大数律，当样本量 n 较大时可以取 EX 的近似 95% 置信区间为 $\bar{X} \pm 2\,\mathrm{SE}(\bar{X})$。 □

例 3.6.2. 考虑线性模型中参数估计的精度。设模型为

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.74}$$

其中 $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I_n)$, σ^2 未知，$\boldsymbol{\beta}$ 是未知系数向量, X 是已知的 $n \times p$ 数值矩阵, $n > p$。在 X 列满秩时 $\boldsymbol{\beta}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}$, 而 $\hat{\boldsymbol{\beta}}$ 的协方差阵为 $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1}$。所以，第 j 个系数 β_j 的标准误差可估计为 $\mathrm{SE}(\hat{\theta}_j) = \hat{\sigma}\sqrt{a^{(jj)}}$, 其中 $\hat{\sigma}$ 是 σ 的估计，$a^{(ij)}$ 为 $(X^T X)^{-1}$ 的 (i,j) 元素。 □

例 3.6.3. 设总体 $X \sim p(x,\theta)$, $\theta \in \Theta$, $X_1, X_2, \ldots, X_n$ 为 X 的简单随机样本，$\hat{\theta}$ 是真值 θ 的最大似然估计。在适当正则性条件下，$\hat{\theta}$ 渐近正态分布，渐近方差为 $\frac{1}{n} I^{-1}(\theta)$, $I(\theta)$ 为参数 θ 的

表 3.1: 百分比的两种置信区间的模拟比较结果 (有删减)

置信度	样本量	p	r_1	r_2	$\bar{l}_1$	$\bar{l}_2$	s_1	s_2
0.99	30	0.1	95.7%	99.2%	0.2646	0.2870	0.0834	0.0448
		0.3	96.8%	99.2%	0.4221	0.3905	0.0375	0.0267
		0.5	98.4%	99.5%	0.4623	0.4196	0.0115	0.0084
	120	0.05	94.2%	99.3%	0.1000	0.1088	0.0204	0.0164
		0.1	98.3%	99.0%	0.1394	0.1423	0.0176	0.0154
		0.3	98.7%	98.8%	0.2144	0.2099	0.0088	0.0080
		0.5	98.7%	99.2%	0.2342	0.2280	0.0014	0.0013
	480	0.01	95.2%	99.0%	0.0227	0.0264	0.0057	0.0045
		0.05	98.1%	99.1%	0.0510	0.0521	0.0049	0.0046
		0.1	98.7%	98.8%	0.0703	0.0707	0.0043	0.0042
		0.3	99.0%	98.9%	0.1076	0.1070	0.0022	0.0021
		0.5	99.1%	99.1%	0.1174	0.1166	0.0002	0.0002
		0.7	99.0%	98.9%	0.1076	0.1070	0.0022	0.0021
		0.9	98.7%	98.8%	0.0703	0.0707	0.0043	0.0042
		0.95	98.1%	99.1%	0.0510	0.0521	0.0049	0.0046
		0.99	95.3%	99.0%	0.0227	0.0264	0.0057	0.0045
0.8	30	0.1	74.3%	88.4%	0.1317	0.1370	0.0414	0.0323
		0.3	75.7%	84.0%	0.2100	0.2058	0.0186	0.0170
		0.5	80.0%	80.0%	0.2300	0.2241	0.0057	0.0053
	120	0.05	77.7%	86.4%	0.0498	0.0510	0.0101	0.0095
		0.1	76.8%	83.1%	0.0693	0.0698	0.0087	0.0084
		0.3	80.6%	80.6%	0.1067	0.1061	0.0044	0.0043
		0.5	76.6%	83.0%	0.1165	0.1157	0.0007	0.0007
	480	0.01	80.4%	74.7%	0.0113	0.0118	0.0028	0.0026
		0.05	78.8%	82.9%	0.0254	0.0255	0.0024	0.0024
		0.1	79.7%	80.4%	0.0350	0.0350	0.0021	0.0021
		0.3	80.3%	78.6%	0.0535	0.0535	0.0011	0.0011
		0.5	81.5%	81.5%	0.0584	0.0583	0.0001	0.0001
		0.7	80.4%	78.7%	0.0535	0.0535	0.0011	0.0011
		0.9	79.8%	80.5%	0.0350	0.0350	0.0021	0.0021
		0.95	78.9%	82.8%	0.0254	0.0255	0.0024	0.0024
		0.99	80.4%	74.6%	0.0113	0.0118	0.0028	0.0026

信息量 (参见茆诗松等 (2006)[7]§2.5.2 定理 2.14):

$$I(\theta)=\mathrm{E}\left[\left(\frac{\partial \ln p(X,\theta)}{\partial\theta}\right)^2\right]=\mathrm{Var}\left(\frac{\partial \ln p(X,\theta)}{\partial\theta}\right) \tag{3.75}$$

在加强的条件下还有

$$I(\theta)=-\mathrm{E}\left(\frac{\partial^2 \ln p(X,\theta)}{\partial\theta^2}\right) \tag{3.76}$$

可以用 $\sqrt{I^{-1}(\hat\theta)/n}$ 估计 $\hat\theta$ 的 SE。 □

例 3.6.4. 设总体 $X\sim p(x,\boldsymbol\theta)$, $\boldsymbol\theta=(\theta_1,\ldots,\theta_m)$，记

$$\boldsymbol S(\theta)=\nabla_{\boldsymbol\theta}\ln p(X,\boldsymbol\theta)=\left(\frac{\partial \ln p(X,\boldsymbol\theta)}{\partial\theta_1},\ldots,\frac{\partial \ln p(X,\boldsymbol\theta)}{\partial\theta_m}\right)^T, \tag{3.77}$$

$$I(\theta)=\mathrm{Var}\left(\boldsymbol S(\theta)\right), \tag{3.78}$$

称 $I(\theta)$ 为信息量矩阵，其 (i,j) 元素为

$$\mathrm{Cov}\left(\frac{\partial \ln p(X,\boldsymbol\theta)}{\partial\theta_i},\frac{\partial \ln p(X,\boldsymbol\theta)}{\partial\theta_j}\right) \tag{3.79}$$

在加强的条件下 $I(\theta)=-E(H(X;\theta))$, H 是 $\ln p(X,\boldsymbol\theta)$ 关于自变量 $\boldsymbol\theta$ 的海色阵，其 (i,j) 元素为 $\frac{\partial^2 \ln p(X,\boldsymbol\theta)}{\partial\theta_i\partial\theta_j}$。设 $X_1,X_2,\ldots,X_n$ 为 X 的简单随机样本，$\hat{\boldsymbol\theta}$ 为 $\boldsymbol\theta$ 的最大似然估计，在适当条件下 $\hat{\boldsymbol\theta}$ 渐近正态分布 $N(\boldsymbol\theta,\frac{1}{n}I^{-1}(\boldsymbol\theta))$, 可以用 $-\frac{1}{n}H^{-1}(X;\hat\theta)$ 作为 $\mathrm{Var}(\hat\theta)$ 的估计。 □

3.6.2 Bootstrap 方法的引入

计算参数估计的标准误差不一定总有简单的公式。例如，需要估计的参数不一定是 EX 这样的简单特征，像中位数、相关系数这样的参数估计的标准误差就比 EX 的估计的标准误差要困难得多。在线性模型估计的例子中，如果独立性、线性或者正态分布的假定不满足则求参数估计方差阵变得很困难，比如稳健回归系数的标准误差就很难得到理论公式。在最大似然估计问题中，最大似然估计不一定总是渐近正态的，信息量有时不存在或难以计算，从而无法用上面的方法给出标准误差。

设总体 X 服从某个未知分布 $F(x)$，$\boldsymbol X=(x_1,x_2,\ldots,x_n)$ 是 X 的一个样本，ϕ 是 F 的一个参数，可以把 ϕ 看成 F 的一个泛函 $\phi(F)$，用统计量 $\hat\phi=g(\boldsymbol X)$ 估计 ϕ，设 $\psi=\psi(g,F,n)$ 是统计量 $\hat\phi$ 的某种分布特征 ($\hat\phi$ 的抽样分布的数字特征)。例如 $\psi=\sqrt{\mathrm{Var}(\bar X)}$ 为统计量 $\bar X$ 的标准误差，又如取 $\psi=E\hat\phi-\phi$ 为统计量 $\hat\phi$ 的偏差。可以用随机模拟的方法估计 ψ。

用随机模拟方法估计 ψ 的步骤如下。

(1) 从样本 $\boldsymbol{X}$ 估计总体分布 F 为 $\hat{F}$;

(2) 从 $\hat{F}$ 抽取 B 个独立样本 $\boldsymbol{Y}^{(b)}$, $b=1,\ldots,B$，每一个 $\boldsymbol{Y}^{(b)}$ 样本量为 n, 称 $\boldsymbol{Y}^{(b)}$ 为 bootstrap 样本。

(3) 从每个 bootstrap 样本 $\boldsymbol{Y}^{(b)}$ 可以估计得到 $\hat{\phi}^{(b)}=g(\boldsymbol{Y}^{(b)})$, $b=1,\ldots,B$。

(4) $\hat{\phi}^{(b)}, b=1,\ldots,B$ 是 $g(\boldsymbol{Y})$ 在 $\hat{F}$ 下的独立同分布样本, 可以用标准的估计方法估计关于 $g(\boldsymbol{Y})$ 在 $\hat{F}$ 下的分布特征 $\hat{\psi}=\psi(g,\hat{F},n)$, 估计结果记作 $\tilde{\psi}$，并以 $\tilde{\psi}$ 作为统计量 $\hat{\phi}$ 的抽样分布的数字特征 $\psi(g,F,n)$ 的估计值。

从样本 $\boldsymbol{X}$ 估计 $\hat{F}$ 时，可以采用参数模型，也可以采用经验分布函数 F_n。参数模型在模型正确时效率较高；经验分布法使用简单，基本不依赖于模型。从经验分布 F_n 抽样，相当于从 $\boldsymbol{X}=(x_1,\ldots,x_n)$ 独立有放回抽样。

估计量的标准误差可以用 bootstrap 方法估计。

例 3.6.5. 设 (H,W) 为某地小学五年级学生的身高和体重的总体，$(H,W)\sim F(\cdot,\cdot)$, 考虑 H 和 W 的相关系数 ϕ 的估计。设调查了 $n=10$ 个学生的身高和体重的数据 $(h_i,w_i), i=1,2,\ldots,n$:

h_i	144	166	163	143	152	169	130	159	160	175
w_i	38	44	41	35	38	51	23	51	46	51

计算得 $\hat{\phi}=g(h_1,w_1,\ldots,h_n,w_n)=0.904$。令 $\mathrm{SE}(\hat{\phi})=\left[\mathrm{Var}(\hat{\phi})\right]^{1/2}=\psi(g,F,n)$。设 $\hat{F}$ 为 F 的估计，取为经验分布 F_n，则 bootstrap 方法用随机模拟方法估计 $\psi(g,F_n,n)$，然后当作 $\mathrm{SE}(\hat{\phi})$ 的估计。计算步骤如下：

(1) 从 F_n 中作 $n=10$ 次独立抽样，即从 $\{(h_1,w_1),\ldots,(h_n,w_n)\}$ 中有放回独立抽取 n 次，得到 $\hat{F}=F_n$ 的一组样本 $\boldsymbol{Y}^{(1)}=((h_1^{(1)},w_1^{(1)}),\ldots,(h_n^{(1)},w_n^{(1)}))$;

(2) 重复第 (1) 步, 直到获取了 B 组 bootstrap 样本 $\boldsymbol{Y}^{(b)}, b=1,\ldots,B$;

(3) 对每一样本 $\boldsymbol{Y}^{(b)}$ 计算样本相关系数 $\hat{\phi}^{(b)}=g(\boldsymbol{Y}^{(b)})$, $b=1,\ldots,B$；

(4) 把 $\hat{\phi}^{(b)}, b=1,\ldots,B$ 作为 $\hat{F}$ 下 $n=10$ 的样本相关系数的简单随机样本，估计其样本标准差 S，以 S 作为 $\psi(g,\hat{F},n)$ 的估计, 进而用 S 估计 $\hat{\phi}$ 在真实的总体分布 F 下的标准误差 $\mathrm{SE}(\hat{\phi})$。

取 $B=10000$ 的一次 bootstrap 计算得到的标准误差估计为 $S=0.101$。当 $B\to\infty$ 时 $S\to\psi(g,F_n,n)$, 但是要注意，由于抽样误差影响，$\psi(g,F_n,n)$ 和 $\psi(g,F,n)$ 之间的误差无法避免。

也可以用参数方法估计 $\hat{F}$，比如从历史经验知道总体的身高、体重服从联合正态分布，就可以按照联合正态总体模型从样本中得到参数最大似然估计后作为 $\hat{F}$ 的参数，这时 $\hat{F}$ 是一个参数确定的二元联合正态分布 $\mathrm{N}(156.1, 41.8, 13.78^2, 8.85^2, 0.904)$。从 $\hat{F}$ 中独立抽样 n 个得到一组样本，共生成 B 组这样的样本，称为 bootstrap 样本。接下来的步骤只要按照上面的 (3)、(4) 估计 $\mathrm{SE}(\hat{\phi})$ 就可以了。取 $B = 10000$ 的一次 bootstrap 计算得到的标准误差估计为 0.080。 □

例 3.6.6. 考虑回归模型系数估计的标准误差计算。一般的回归模型可以写成

$$y_i = h(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{3.80}$$

其中 h 已知，$\boldsymbol{\beta}$ 是未知参数向量，$\{\varepsilon_i\}$ iid $F(x)$，$F(x)$ 未知, $\{\boldsymbol{x}_i\}$ 是确定数值向量。可以用最小二乘等方法得到 $\boldsymbol{\beta}$ 的估计 $\hat{\boldsymbol{\beta}} = g(y_1, \ldots, y_n)$, 希望估计参数估计的协方差阵 $\mathrm{Var}(\hat{\boldsymbol{\beta}})$, 协方差阵主对角线元素的平方根就是单个系数估计的标准误差。这个模型中，未知的分布信息包括 $\boldsymbol{\beta}$ 和 F。$\boldsymbol{\beta}$ 可用 $\hat{\boldsymbol{\beta}}$ 估计, F 可以用回归残差的经验分布来估计或假设一个参数模型估计模型参数。

用 bootstrap 方法估计 $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ 的步骤如下:

(1) 估计 $\boldsymbol{\beta}$ 得到 $\hat{\boldsymbol{\beta}} = g(y_1, \ldots, y_n)$;

(2) 计算残差 $e_i = y_i - h(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})$, $i = 1, \ldots, n$;

(3) 对 $b = 1, \ldots, B$ 重复：从 $\{e_1, \ldots, e_n\}$ 中有放回独立抽取 n 次得 $\{e_i^{(b)}, i = 1, \ldots, n\}$;

(4) 令 $y_i^{(b)} = h(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}) + e_i^{(b)}$, $i = 1, \ldots, n$, $b = 1, 2, \ldots, B$;

(5) 对 $b = 1, 2, \ldots, B$ 重复：从 $(y_1^{(b)}, \ldots, y_n^{(b)})$ 中估计 $\hat{\boldsymbol{\beta}}^{(b)} = g(y_1^{(b)}, \ldots, y_n^{(b)})$。

(6) 用 $\hat{\boldsymbol{\beta}}^{(b)}, b = 1, \ldots, B$ 的样本方差阵估计 $\mathrm{Var}(\hat{\boldsymbol{\beta}})$。

□

3.6.3 Bootstrap 偏差校正

设 $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ 为总体 $F(\cdot)$ 的样本，总体参数 $\phi = \phi(F)$ 的估计为 $\hat{\phi} = g(\boldsymbol{X})$, $b = E\hat{\phi} - \phi$ 为估计偏差，$\mathrm{Var}(\hat{\phi})$ 为估计方差，估计的均方误差可以分解为

$$E[\hat{\phi} - \phi]^2 = \mathrm{Var}(\hat{\phi}) + b^2. \tag{3.81}$$

如果 $\hat{b}$ 是 b 的估计，则参数 ϕ 的一个改善的估计为 $\tilde{\phi}=\hat{\phi}-\hat{b}$，新的估计在减小了偏差的同时一般也减小了均方误差。设 $b=\psi(g,F,n)$, $\hat{F}$ 是总体分布 F 的一个估计，这里 $\hat{F}$ 取为经验分布 F_n，则可以用 $\hat{b}=\psi(g,\hat{F},n)=Eg(\boldsymbol{Y})-\hat{\phi}$ 来估计偏差, 其中 $\boldsymbol{Y}$ 是总体 F_n 的样本量为 n 的样本, $\hat{\phi}$ 恰好是总体分布为 F_n 时的参数 ϕ, 即 $\hat{\phi}=\phi(F_n)$。如果 $\hat{b}$ 不能通过理论公式计算，可以用 bootstrap 方法估计 $\hat{b}$，步骤如下:

(1) 从 $\{x_1,x_2,\ldots,x_n\}$ 独立有放回地抽取 n 个，记为 $\boldsymbol{Y}^{(1)}=(Y_1^{(1)},Y_2^{(1)},\ldots,Y_n^{(1)})$。

(2) 重复第 (1) 步，直到获取了 B 组 bootstrap 样本 $\boldsymbol{Y}^{(b)},b=1,2,\ldots,B$;

(3) 从每个 bootstrap 样本 $\boldsymbol{Y}^{(b)}$ 可以估计得到 $\hat{\phi}^{(b)}=g(\boldsymbol{Y}^{(b)}),\ b=1,\ldots,B$。

(4) 用 $\hat{\phi}^{(b)},b=1,\ldots,B$ 作为 $g(\boldsymbol{Y})$ 在 F_n 下的独立同分布样本, 估计 $\hat{b}=\psi(g,F_n,n)$ 为 $\tilde{b}=\frac{1}{B}\sum_{b=1}^{B}\hat{\phi}^{(b)}-\hat{\phi}$。

最后，取 $\tilde{\phi}=\hat{\phi}-\tilde{b}=2\hat{\phi}-\frac{1}{B}\sum_{b=1}^{B}\hat{\phi}^{(b)}$ 为改善的估计。

例 3.6.7. 设 $X\sim \mathrm{N}(\mu,\sigma^2)$, μ,σ^2 未知，$x_1,x_2,\ldots,x_n$ 为 X 的样本。考虑 $\phi=\mu^2$ 的估计。用最大似然估计法估计 ϕ 为 $\hat{\phi}=\bar{X}^2$，其中 $\bar{X}$ 为样本平均值。令 $Z=\sqrt{n}(\bar{X}-\mu)/\sigma$, 则 $Z\sim \mathrm{N}(0,1)$。可以计算出估计偏差为

$$b=E\bar{X}^2-\mu^2=E(\mu+\frac{\sigma}{\sqrt{n}}Z)^2-\mu^2=\frac{\sigma^2}{n}, \tag{3.82}$$

估计的均方误差为

$$\begin{aligned} L_0=&E(\bar{X}^2-\mu^2)^2=E\left(\frac{2\sigma\mu}{\sqrt{n}}Z+\frac{\sigma^2}{n}Z^2\right)^2\\ =&\frac{4\sigma^2\mu^2}{n}+\frac{3\sigma^4}{n^2}. \end{aligned} \tag{3.83}$$

估计 b 为 $\hat{b}_1=S^2/n$(S^2 为样本方差)，用 $\hat{\phi}_1=\bar{X}^2-\hat{b}_1=\bar{X}^2-\frac{S^2}{n}$ 作为 μ^2 的改善的估计, 则 $\hat{\phi}_1$ 的均方误差比 $\hat{\phi}$ 的均方误差减小了 $\frac{n-3}{n-1}\frac{\sigma^4}{n^2}$ (设 $n>3$, 见习题17)。

如果模型更为复杂，比如，总体分布类型未知，$\hat{b}_1$ 这样的简单偏差估计很难得到，这种情况下可以用 bootstrap 方法进行偏差校正，步骤如下：

(1) 对 $b=1,2,\ldots,B$ 重复: 从 $x_1,x_2,\ldots,x_n$ 独立有放回地抽取 n 个，组成 bootstrap 样本 $\boldsymbol{Y}^{(b)}=(y_1^{(b)},\ldots,y_n^{(b)})$;

(2) 对每个 bootstrap 样本计算 $\hat{\phi}^{(b)}=\left(\frac{1}{n}\sum_{i=1}^{n}y_i^{(b)}\right)^2$;

(3) 用 $\tilde{\phi} = 2\bar{X}^2 - \frac{1}{B}\sum_{b=1}^{B}\hat{\phi}^{(b)}$ 作为 μ^2 的改善的估计。

□

Jacknife 方法是另外一种对估计量的偏差和方差进行估计的方法，这种方法不需要从原来的样本重新随机抽样，而是把原来的 n 个样本点分为 r 份，每次删去其中一份后计算统计量值，利用 r 个这样的统计量值对估计量的偏差和方差进行估计。详见 Gentle(2002)[18]§3.3。

3.6.4 Bootstrap 置信区间

枢轴量法是构造置信区间的最基本的方法。设 ϕ 是总体 $F(\cdot)$ 的一个参数, 看成 F 的一个泛函 $\phi = \phi(F)$。$\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ 为总体的样本，$g(\boldsymbol{X})$ 为与 ϕ 有关系的一个统计量，经常是 ϕ 的估计量。如果有变换 $W = h(g(\boldsymbol{X}), \phi)$ 使得 W 的分布不依赖于任何未知参数，则设 W 的左右两侧分位数分别为 $w_{\frac{\alpha}{2}}$ 和 $w_{1-\frac{\alpha}{2}}$, 有

$$P(w_{\frac{\alpha}{2}} < h(T, \phi) < w_{1-\frac{\alpha}{2}}) = 1 - \alpha, \tag{3.84}$$

反解上面的不等式可以得到 ϕ 的置信区间。

如果对枢轴量 W 很难求分位数时，可以用 bootstrap 方法获得置信区间。设 $\hat{F}$ 为总体分布 F 的估计，设 $\boldsymbol{Y} = (y_1, \ldots, y_n)$ 为总体 $\hat{F}$ 的样本，$\hat{\phi} = \phi(\hat{F})$ 为与总体 $\hat{F}$ 对应的参数 ϕ 的值，实际是 ϕ 的估计值，则 $V = h(g(\boldsymbol{Y}), \hat{\phi})$ 与 W 的分布相近，可以用 V 的分位数近似 W 的分位数。

例 3.6.8. 设总体 $X \sim F(x, \theta)$, θ 为总体的未知参数，$\phi = EX$, $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ 为总体的样本，则 $g(\boldsymbol{X}) = \bar{X}$ 是 ϕ 的估计，若 $W = h(g(\boldsymbol{X}), \phi) = \bar{X} - EX$ 的分布与 θ 无关，求 W 的分位数 $w_{\frac{\alpha}{2}}$ 和 $w_{1-\frac{\alpha}{2}}$ 就可以构造 $\phi = EX$ 的置信区间 $(\bar{X} - w_{1-\frac{\alpha}{2}},\ \bar{X} - w_{\frac{\alpha}{2}})$。

若 W 的分位数无法求得，用经验分布 F_n 作为总体分布 F 的估计，这时 $\phi(F_n) = \bar{X}$, 设 $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ 为 F_n 的样本，$V = h(g(\boldsymbol{Y}), \bar{X}) = \bar{Y} - \bar{X}$，这里 $\bar{X}$ 作为已知值，可以用 V 的分位数近似代替 W 的分位数。求 V 的分位数，只要用有放回独立抽样方法从 $x_1, x_2, \ldots, x_n$ 抽取 F_n 的 B 组样本 $\boldsymbol{Y}^{(b)} = (y_1^{(b)}, \ldots, y_n^{(b)}), b = 1, 2, \ldots, B$, 对每组样本计算平均值 $\bar{Y}^{(b)}$, 定义 $V^{(b)} = \bar{Y}^{(b)} - \bar{X}$, 用 $V^{(b)}, b = 1, 2, \ldots, B$ 的样本分位数估计 V 的分位数，作为 W 的分位数 $w_{\frac{\alpha}{2}}$ 和 $w_{1-\frac{\alpha}{2}}$ 的近似。 □

3.7 MCMC

3.7.1 马氏链和 MCMC 介绍

实际工作中经常遇到分布复杂的高维随机向量抽样问题。§3.2.4的重要抽样法可以应付维数不太高的情况，但是对于维数很高而且分布很复杂（比如，分布密度多峰而且位置不易确定的情况）则难以处理。

MCMC(马氏链蒙特卡洛) 是一种对高维随机向量抽样的方法，此方法模拟一个马氏链，使马氏链的平稳分布为目标分布，由此产生大量的近似服从目标分布的样本，但样本不是相互独立的。MCMC 的目标分布密度函数或概率函数可以只计算到差一个常数倍的值。MCMC 方法适用范围广，近年来获得了广泛的应用。

先介绍马氏链的概念。

设 $\{X_t, t=0,1,\dots\}$ 为随机变量序列，称为一个**随机过程**。称 X_t 为“系统在时刻 t 的**状态**”。为讨论简单起见，设所有 X_t 均取值于有限集合 $S=\{1,2,\dots,m\}$, 称 S 为**状态空间**。如果 $\{X_t\}$ 满足

$$\begin{aligned} &P(X_{t+1}=j|X_0=k_0,\dots X_{t-1}=k_{t-1},X_t=i)\\ =&P(X_{t+1}=j|X_t=i)=p_{ij},\ t=0,1,\dots,\ k_0,\dots,k_{t-1},i,j\in S, \end{aligned} \tag{3.85}$$

则称 $\{X_t\}$ 为**马氏链**, p_{ij} 为**转移概率**，矩阵 $P=(p_{ij})_{m\times m}$ 为**转移概率矩阵**。显然 $\sum_{j=1}^m p_{ij}=1,\ i=1,2,\dots,m$。对马氏链，$P(X_{t+k}=j|X_t=i)\triangleq p_{ij}^{(k)}$ 也不依赖于 t，称为k **步转移概率**。如果对任意 $i,j\in S, i\neq j$ 都存在 $k\geq 1$ 使得 $p_{ij}^{(k)}>0$ 则称 $\{X_t\}$ 为**不可约马氏链**。不可约马氏链的所有状态是互相连通的，即总能经过若干步后互相转移。对马氏链 $\{X_t\}$ 的某个状态 i, 如果存在 $k\geq 0$ 使得 $p_{ii}^{(k)}>0$ 并且 $p_{ii}^{(k+1)}>0$，则称 i 是**非周期**的。如果一个马氏链所有状态都是非周期的，则该马氏链称为非周期的。不可约马氏链只要有一个状态是非周期的则所有状态是非周期的。对只有有限个状态的非周期不可约马氏链有

$$\lim_{n\to\infty} P(X_n=j)=\pi_j,\ j=1,2,\dots,m, \tag{3.86}$$

其中 $\{\pi_j, j=1,2,\dots,m\}$ 为常数, 称为 $\{X_t\}$ 的**极限分布**。$\{\pi_j\}$ 满足方程组

$$\begin{cases} \sum_{i=1}^m \pi_i p_{ij}=\pi_j,\ j=1,2,\dots,m\\ \sum_{j=1}^m \pi_j=1, \end{cases} \tag{3.87}$$

称满足(3.87)的分布 $\{\pi_j\}$ 为**平稳分布**。对只有有限个状态的非周期不可约马氏链，极限分布和平稳分布存在且为同一分布。

如果允许状态空间 S 为可列个元素，极限分布的条件需要更多的讨论。对状态 i，如从状态 i 出发总能再返回状态 i，则称状态 i 是**常返**的 (recurrent)。对常返状态 i，如果从 i 出发首次返回 i 的时间的期望有限，称 i 是**正常返**的。非周期正常返状态称为**遍历**的。所有状态都遍历的马氏链称为遍历马氏链。非周期遍历马氏链存在唯一的极限分布和平稳分布，且二者相同。

如果存在 $\{\pi_j, j = 1, 2, \ldots, m\}$, $\pi_j \geq 0$, $\sum_{j=1}^{m} \pi_j = 1$, 使得

$$\pi_i p_{ij} = \pi_j p_{ji}, \ \forall i \neq j, \tag{3.88}$$

称这样的马氏链为**细致平衡**的 (detailed balanced)，这时 $\{\pi_j\}$ 是 $\{X_t\}$ 的平稳分布。事实上，若 $P(X_t = i) = \pi_i, i \in S$, 则

$$P(X_{t+1} = j) = \sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j, \ \forall j \in S. \tag{3.89}$$

马氏链的概念可以推广到 X_t 的取值集合 $\mathcal{X}$ 为可列集或 $\mathbb{R}^d$ 的区域的情形。如果各 $\{X_t\}$ 的有限维分布是连续型的，则(3.85)可以改用条件密度表示, 这时的 $\{X_t\}$ 按照随机过程论中的习惯应该称作马氏过程, 但这里还是叫做马氏链。

如果遍历的不可约马氏链 $\{X_t\}$ 有平稳分布 $\pi(x), x \in \mathcal{X}$, 则从任意初值出发模拟产生序列 $\{X_t\}$，当 t 很大时，X_t 的分布就近似服从 π，抛弃开始的一段后的 X_t 序列可以作为分布 π 的相关的样本，抛弃的一段序列叫做**老化期**。设 $Y \sim \pi(\cdot)$, $h(y), y \in \mathcal{X}$ 是有界函数, 为估计 $\theta \triangleq Eh(Y)$，用 $X_t, t = k+1, \ldots, n$ 作为 π 的样本，用估计量 $\hat{\theta} = \frac{1}{n-k} \sum_{t=k+1}^{n} h(X_t)$ 来估计 θ，则 $\hat{\theta}$ 是 θ 的强相合估计。老化期长度 k 可以从 $\hat{\theta}$ 的变化图形经验地选取。

这样的估计量 $\hat{\theta}$ 是相关样本的平均值，无法用原来独立样本的公式估计 $\mathrm{Var}(\hat{\theta})$ 从而得到 $\hat{\theta}$ 的标准误差。为了估计 $\mathrm{Var}(\hat{\theta})$, 可以采用如下的分段平均法。把样本 $X_{k+1}, \ldots, X_n$ 分为 s 段，每段 r 个 (设 $n - k = sr$)。设第 j 段的 r 个 $h(X_t)$ 的平均值为 $Z_j, j = 1, 2, \ldots, s$, 设 $\{Z_j, j = 1, 2, \ldots, s\}$ 的样本方差为 $\hat{\sigma}^2 = \frac{1}{s-1} \sum_{j=1}^{s} (Z_j - \bar{Z})^2$，因为 $\hat{\theta}$ 等于 $\{Z_j, j = 1, 2, \ldots, s\}$ 的样本均值 $\bar{Z}$，当 r 足够大时，可以认为各 Z_j 相关性已经很弱，这时 $\hat{\theta}$ 的方差可以用 $\hat{\sigma}^2/s$ 估计。r 的大小依赖于不同时刻的 X_t 的相关性强弱，相关性越强，需要的 r 越大。

以上的方法就是 MCMC 方法（马氏链蒙特卡洛）。一般地，对高维或取值空间 $\mathcal{X}$ 结构复杂的随机向量 X, MCMC 方法构造取值于 $\mathcal{X}$ 的马氏链，使其平稳分布为 X 的目标分布。模拟此马氏链，抛弃开始的部分抽样值，把剩余部分作为 X 的非独立抽样。非独立抽样的估计效率比独立抽样低。

MCMC 方法的关键在于如何从第 t 时刻转移到第 $t+1$ 时刻。好的转移算法应该使得马氏链比较快地收敛到平稳分布，并且不会长时间地停留在取值空间 $\mathcal{X}$ 的局部区域内（在目标分布是多峰分布且峰高度差异较大时容易出现这种问题）。

Metropolis-Hasting 方法 (MH 方法) 是一个基本的 MCMC 算法，此算法在每一步试探地进行转移 (如随机游动)，如果转移后能提高状态 x_t 在目标分布 π 中的密度值则接受转移结果，否则以一定的概率决定是转移还是停留不动。

Gibbs 抽样是另外一种常用的 MCMC 方法，此方法轮流延各个坐标轴方向转移，且转移概率由当前状态下用其它坐标预测转移方向坐标的条件分布给出。因为利用了目标分布的条件分布，所以 Gibbs 抽样方法的效率比 MH 方法效率更高。

3.7.2 Metropolis-Hasting 抽样

设随机变量 X 分布为 $\pi(x), x \in \mathcal{X}$。为论述简单起见仍假设 $\mathcal{X}$ 是离散集合。算法需要一个试转移概率函数 $T(y|x), x, y \in \mathcal{X}$，满足 $0 \le T(y|x) \le 1, \sum_y T(y|x) = 1$，并且

$$T(y|x) > 0 \Leftrightarrow T(x|y) > 0. \tag{3.90}$$

算法首先从 $\mathcal{X}$ 中任意取初值 $X^{(0)}$。设经过 t 步后算法的当前状态为 $X^{(t)}$, 则下一步由试转移分布 $T(y|X^{(t)})$ 抽取 Y，并生成 $U \sim$U(0,1), 然后按如下规则转移:

$$X^{(t+1)} = \begin{cases} Y & \text{若} U \le r(X^{(t)}, Y) \\ X^{(t)} & \text{否则} \end{cases} \tag{3.91}$$

其中

$$r(x, y) = \min\left\{1, \frac{\pi(y)T(x|y)}{\pi(x)T(y|x)}\right\}. \tag{3.92}$$

在 MH 算法中如果取 $T(y|x) = T(x|y)$，则 $r(x, y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right)$, 相应的算法称为 Metropolis **抽样法**。

如果取 $T(y|x) = g(y)$(不依赖于 x)，则 $r(x, y) = \min\left(1, \frac{\pi(y)/g(y)}{\pi(x)/g(x)}\right)$, 相应的算法称为 Metropolis **独立抽样法**, 和重要抽样有相似之处, 试抽样分布 $g(\cdot)$ 经常取为相对重尾的分布。

在 MH 算法中，目标分布 $\pi(x)$ 可以用差一个常数倍的 $\tilde{\pi}(x) = C\pi(x)$ 代替，这样关于目标分布仅知道差一个常数倍的 $\tilde{\pi}(x)$ 的情形，也可以使用此算法。

下面说明 MH 抽样方法的合理性。我们来验证 MH 抽样的转移概率 $A(x, y) = P(X^{(t+1)} = y|X^{(t)} = x)$ 满足细致平衡条件。易见

$$A(x, y) = \begin{cases} T(y|x)r(x, y), & y \ne x, \\ T(x|x) + \sum_{z \ne x} T(z|x)[1 - r(x, z)], & y = x, \end{cases} \tag{3.93}$$

于是当 $x \neq y$ 时

$$\pi(x)A(x,y) = \pi(x)T(y|x)\min\left\{1, \frac{\pi(y)T(x|y)}{\pi(x)T(y|x)}\right\} = \min\left\{\pi(x)T(y|x), \pi(y)T(x|y)\right\},$$

等式右侧关于 x, y 是对称的，所以等式左侧把 x, y 交换后仍相等。所以，MH 构造的马氏链以 $\{\pi(x)\}$ 为平稳分布。多数情况下 MH 构造的马氏链也以 $\{\pi(x)\}$ 为极限分布。

(3.92)中的 $r(x,y)$ 还可以推广为如下的形式

$$\tilde{r}(x,y) = \frac{\alpha(x,y)}{\pi(x)T(y|x)}, \tag{3.94}$$

其中 $\alpha(x,y)$ 是任意的满足 $\alpha(x,y) = \alpha(y,x)$ 且使得 $\tilde{r}(x,y) \leq 1$ 的函数。易见这样的 $\tilde{r}(x,y)$ 仍使得生成的马氏链满足细致平衡条件。

例 3.7.1. X 的取值集合 $\mathcal{X}$ 可能是很大的，以至于无法穷举，目标分布 $\pi(x)$ 可能是只能确定到差一个常数倍。

例如，设

$$\mathcal{X} = \{\boldsymbol{x} = (x_1, x_2, \ldots, x_n) : (x_1, x_2, \ldots, x_n)\text{为}(1, 2, \ldots, n)\text{的一个排列, 并满足}\sum_{j=1}^{n} j x_j > a\},$$

其中 a 是一个给定的常数。用 $|\mathcal{X}|$ 表示 $\mathcal{X}$ 的元素个数，当 n 较大时 $\mathcal{X}$ 是 $(1, 2, \ldots, n)$ 的所有 $n!$ 个排列的一个子集，$|\mathcal{X}|$ 很大，很难穷举 $\mathcal{X}$ 的元素, 从而 $|\mathcal{X}|$ 未知。

设 X 服从 $\mathcal{X}$ 上的均匀分布，即 $\pi(\boldsymbol{x}) = C,\ x \in \mathcal{X}, C = 1/|\mathcal{X}|$ 但 C 未知。要用 MH 方法产生 X 的抽样序列。

试抽样 $T(\boldsymbol{y}|\boldsymbol{x})$ 如果允许转移到所有的 $\boldsymbol{y}$ 是很难执行的，因为 $\boldsymbol{y}$ 的个数太多了。我们定义一个 $\boldsymbol{x}$ 的近邻的概念，仅考虑转移到 $\boldsymbol{x}$ 的近邻。一种定义是，如果把 $\boldsymbol{x}$ 的 n 个元素中的某两个交换位置后可以得到 $\boldsymbol{y} \in \mathcal{X}$，则 $\boldsymbol{y}$ 称为 $\boldsymbol{x}$ 的一个近邻，记 $N(\boldsymbol{x})$ 为 $\boldsymbol{x}$ 的所有近邻的集合，记 $|N(\boldsymbol{x})|$ 为 $\boldsymbol{x}$ 的近邻的个数。当 n 很大时，求 $N(\boldsymbol{x})$ 也需要从 $C_n^2 = \frac{1}{2}n(n-1)$ 个可能的元素中用穷举法选择。取试转移概率函数为

$$T(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{|N(\boldsymbol{x})|},\ \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X},$$

即从 $\boldsymbol{x}$ 出发，等可能地试转移到 $\boldsymbol{x}$ 的任何一个近邻上。

因为目标分布 $\pi(\boldsymbol{x})$ 是常数，所以这时

$$r(\boldsymbol{x}, \boldsymbol{y}) = \min\left(1, \frac{|N(\boldsymbol{x})|}{|N(\boldsymbol{y})|}\right),$$

即从 $\boldsymbol{x}$ 试转移到 $\boldsymbol{y}$ 后，如果 $\boldsymbol{y}$ 的近邻数不超过 $\boldsymbol{x}$ 的近邻数则确定转移到 $\boldsymbol{y}$，否则，仅按概率 $|N(\boldsymbol{x})|/|N(\boldsymbol{y})|$ 转移到 $\boldsymbol{y}$。这就构成了对 X 抽样的 MH 算法。 □

连续型分布的 MH 抽样法 对于连续型的目标分布, 设 $\pi(x)$ 为目标分布的密度，这时 $T(y|x)$ 改为给定 x 条件下的试抽样密度，$r(x,y)$ 定义不变，算法和离散型目标分布的情形相同。

例 3.7.2. 考虑一个贝叶斯推断问题。在金融投资中，投资者经常把若干种证券组合在一起来减少风险。假设有 5 支股票的 $n = 250$ 个交易日的收益率记录，每个交易日都找出这 5 支股票收益率最高的一个，设 X_i 表示第 i 支股票在 n 个交易日中收益率为最高的次数 $(i = 1, 2, \ldots, 5)$。设 $(X_1, \ldots, X_5)$ 服从多项分布，相应的概率假设为

$$\boldsymbol{p} = \left(\frac{1}{3}, \frac{1-\beta}{3}, \frac{1-2\beta}{3}, \frac{2\beta}{3}, \frac{\beta}{3}\right),$$

其中 $\beta \in (0, 0.5)$ 为未知参数。假设 β 有先验分布 $p_0(\beta) \sim \mathrm{U}(0, 0.5)$。设 $(x_1, \ldots, x_5)$ 为 $(X_1, \ldots, X_5)$ 的观测值，则 β 的后验分布为

$$\begin{aligned}
&f(\beta|x_1, \ldots, x_5) \propto p(x_1, \ldots, x_5|\beta)p_0(\beta) \\
=&\binom{n}{x_1, \ldots, x_5}\left(\frac{1}{3}\right)^{x_1}\left(\frac{1-\beta}{3}\right)^{x_2}\left(\frac{1-2\beta}{3}\right)^{x_3}\left(\frac{2\beta}{3}\right)^{x_4}\left(\frac{\beta}{3}\right)^{x_5}\frac{1}{0.5}I_{(0,0.5)}(\beta) \\
\propto&(1-\beta)^{x_2}(1-2\beta)^{x_3}\beta^{x_4+x_5}I_{(0,0.5)}(\beta) \triangleq \tilde{\pi}(\beta).
\end{aligned}$$

为了求 β 后验均值，需要产生服从 $f(\beta|x_1, \ldots, x_5)$ 的抽样。从 β 的后验分布很难直接抽样，采用 Metropolis 抽样法。设当前 β 的状态为 $\beta^{(t)}$, 取试抽样分布 $T(y|\beta^{(t)})$ 为 U(0,0.5), 则 $T(y|x) = T(x|y)$,

$$r(\beta^{(t)}, y) = \min\left(1, \frac{\tilde{\pi}(y)}{\tilde{\pi}(\beta^{(t)})}\right) = \min\left(1, \left(\frac{1-y}{1-\beta^{(t)}}\right)^{x_2}\left(\frac{1-2y}{1-2\beta^{(t)}}\right)^{x_3}\left(\frac{y}{\beta^{(t)}}\right)^{x_4+x_5}\right),$$

从 U(0, 0.5) 试抽取 y，以概率 $r(\beta^{(t)}, y)$ 接受 $\beta^{(t+1)} = y$ 即可。 □

随机游动 MH 算法 MH 抽样中试转移概率函数 $T(y|x)$ 较难找到，容易想到的是从 $x^{(t)}$ 作随机游动的试转移方法，叫做随机游动 Metropolis 抽样。

设 X 的目标分布 $\pi(\boldsymbol{x})$ 取值于欧式空间 $\mathcal{X} = \mathbb{R}^d$。从 $\boldsymbol{x}^{(t)}$ 出发试转移，令

$$\boldsymbol{y} = \boldsymbol{x}^{(t)} + \boldsymbol{\varepsilon}_t, \tag{3.95}$$

其中 $\boldsymbol{\varepsilon}_t \sim g(\boldsymbol{x}; \sigma)$ 对不同 t 是独立同分布的, $T(\boldsymbol{y}|\boldsymbol{x}) = g(\boldsymbol{y} - \boldsymbol{x})$。设 g 是关于 $\boldsymbol{x} = \boldsymbol{0}$ 对称的分布，则 $T(\boldsymbol{y}|\boldsymbol{x}) = T(\boldsymbol{x}|\boldsymbol{y})$。

常取 g 为 $\mathrm{N}(\boldsymbol{0}, \sigma^2 I)$ 和半径为 σ 的中心为 $\boldsymbol{0}$ 的球内的均匀分布。

转移法则为: 从 $\boldsymbol{x}^{(t)}$ 出发试转移到 $\boldsymbol{y}$ 后，若 $\pi(\boldsymbol{y}) > \pi(\boldsymbol{x}^{(t)})$ 则令 $\boldsymbol{x}^{(t+1)} = \boldsymbol{y}$; 否则, 独立地抽取 $U \sim \mathrm{U}(0,1)$, 取

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \boldsymbol{y}, & \text{当} U \le \pi(\boldsymbol{y})/\pi(\boldsymbol{x}^{(t)}), \\ \boldsymbol{x}^{(t)}, & \text{否则.} \end{cases}$$

随机游动 MH 算法是一种 Metropolis 抽样方法。随机游动的步幅 σ 是重要参数，步幅过大导致拒绝率大，步幅过小使得序列的相关性太强，收敛到平衡态速度太慢。一个建议选法是试验各种选法，使得试抽样被接受的概率在 0.25 到 0.35 之间。

例 3.7.3. 考虑如下的简单气体模型: 在平面区域 $G = [0, A] \times [0, B]$ 内有 K 个直径为 d 的刚性圆盘。随机向量 $\boldsymbol{X} = (x_1, y_1, \ldots, x_K, y_k)$ 为这些圆盘的位置坐标。分布 $\pi(\boldsymbol{x})$ 是 G 内所有允许位置的均匀分布。希望对 π 抽样。

先找一个初始的允许位置 $\boldsymbol{x}^{(0)}$。比如，把圆盘整齐地排列在左上角。

设已得到 $\boldsymbol{x}^{(t)}$，随机选取一个圆盘 i，把圆盘 i 的位置试移动到 $(x_i', y_i') = (x_i + \delta_i, y_i + \epsilon_i)$，其中 δ_i, ϵ_i 独立同 $\mathrm{N}(0, \sigma^2)$ 分布。如果得到的位置是允许的则接受结果，否则留在原地不动。 □

3.7.3 Gibbs 抽样

一般的 MH 抽样每一步首先进行尝试运动，然后根据新的状态是否靠近目标分布来接受或拒绝试抽样点，所以可能会存在多次的无效尝试，效率较低。

Gibbs 抽样是另外一种 MCMC 方法，它仅在坐标轴方向尝试转移，用当前点的条件分布决定下一步的试抽样分布，所有试抽样都被接受，不需要拒绝，所以效率可以更高。

设状态用 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ 表示, 设目标分布为 $\pi(\boldsymbol{x})$, 用 $\boldsymbol{x}_{(-i)}$ 表示 $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$，假设 $\pi(\cdot)$ 的条件分布 $p(x_i|\boldsymbol{x}_{(-i)})$ 都能够比较容易地抽样。

Gibbs 抽样每一步从条件分布中抽样，可以轮流从每一分量抽样，这样的算法称为系统扫描 Gibbs 抽样算法:

从 $\pi(\boldsymbol{x})$ 的取值区域中任意取一个初值 $\boldsymbol{X}^{(0)}$
for (t **in** 0:$(N-1)$) {
  for (i **in** 1:n) {
    从条件分布 $p(x_i|X_1^*, \ldots, X_{i-1}^*, X_{i+1}^{(t)}, \ldots, X_n^{(t)})$ 中抽取 X_i^*
  }
  令 $\boldsymbol{X}^{(t+1)} \leftarrow (X_1^*, \ldots, X_n^*)$
}

从条件分布抽样的次序也可以是随机选取各个分量，这样的算法称为随机扫描 Gibbs 抽样算法:

从 $\pi(\boldsymbol{x})$ 的取值区域中任意取一个初值 $\boldsymbol{X}^{(0)}$
for (t **in** 0:$(N-1)$) {
按概率 $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_n)$ 从 $(1,\ldots,n)$ 中随机抽取下标 i
从条件分布 $p(x_i|\boldsymbol{X}^{(t)}_{(-i)})$ 中抽取 X_i^*, 令 $\boldsymbol{X}^{(t+1)}=(X_1^{(t)},\ldots,X_{i-1}^{(t)},X_i^*,X_{i+1}^{(t)},\ldots,X_n^{(t)})$
}

其中下标的抽样概率 $\boldsymbol{\alpha}$ 为事先给定。

容易看出，无论采用系统扫描还是随机扫描的 Gibbs 抽样，如果 $\boldsymbol{X}^{(t)}$ 服从目标分布，则 $\boldsymbol{X}^{(t+1)}$ 也服从目标分布。以系统扫描方法为例，设在第 $t+1$ 步已经抽取了 $X_1^*,\ldots,X_{i-1}^*$, 令 $\boldsymbol{Y}=(X_1^*,\ldots,X_{i-1}^*,X_i^{(t)},\ldots,X_n^{(t)})$, 设 $\boldsymbol{Y}\sim\pi(\cdot)$。下一步从 $\pi(\cdot)$ 的边缘密度 $p(x_i|X_1^*,\ldots,X_{i-1}^*,X_{i+1}^{(t)},\ldots,X_n^{(t)})$ 抽取 X_i^*, 则 $\boldsymbol{Y}^*\triangleq(X_1^*,\ldots,X_{i-1}^*,X_i^*,X_{i+1}^{(t)},\ldots,X_n^{(t)})$ 的分布密度在 $\boldsymbol{Y}^*$ 处的值为

$$\begin{aligned}&p(X_1^*,\ldots,X_{i-1}^*,X_i^*,X_{i+1}^{(t)},\ldots,X_n^{(t)})\\=&p(X_i^*|X_1^*,\ldots,X_{i-1}^*,X_{i+1}^{(t)},\ldots,X_n^{(t)})p(X_1^*,\ldots,X_{i-1}^*,X_{i+1}^{(t)},\ldots,X_n^{(t)})\\=&\pi(X_1^*,\ldots,X_{i-1}^*,X_i^*,X_{i+1}^{(t)},\ldots,X_n^{(t)})\end{aligned}$$

即 $\boldsymbol{Y}\sim\pi(\cdot)$ 则 $\boldsymbol{Y}^*\sim\pi(\cdot)$。

例 3.7.4. 设目标分布为二元正态分布，设 $\boldsymbol{X}\sim\pi(\boldsymbol{x})$ 为

$$\mathrm{N}\left\{\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix}\right\}$$

采用系统扫描 Gibbs 抽样方案，每一步的迭代为，

$$\begin{aligned}&\text{抽取}X_1^{(t+1)}|X_2^{(t)}\sim\mathrm{N}(\rho X_2^{(t)},1-\rho^2)\\&\text{抽取}X_2^{(t+1)}|X_1^{(t+1)}\sim\mathrm{N}(\rho X_1^{(t+1)},1-\rho^2)\end{aligned}$$

递推可得

$$\begin{pmatrix}X_1^{(t)}\\X_2^{(t)}\end{pmatrix}\sim\mathrm{N}\left\{\begin{pmatrix}\rho^{2t-1}X_2^{(0)}\\\rho^{2t}X_2^{(0)}\end{pmatrix},\begin{pmatrix}1-\rho^{4t-2}&\rho-\rho^{4t-1}\\\rho-\rho^{4t-1}&1-\rho^{4t}\end{pmatrix}\right\}\tag{3.96}$$

当 $t\to\infty$ 时，$(X_1^{(t)},X_2^{(t)})$ 的期望与目标分布期望之差为 $O(|\rho|^{2t})$, 方差与目标分布方差之差为 $O(|\rho|^{4t})$。 □

例 3.7.5. 设目标分布为

$$\pi(x,y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1},\ x=0,1,\ldots,n,\ 0\le y\le 1, \tag{3.97}$$

则 $X|Y \sim \mathrm{B}(n,y)$, $Y|X \sim \mathrm{Beta}(x+\alpha, n-x+\beta)$。易见 Y 的边缘分布为 $\mathrm{Beta}(\alpha,\beta)$。可以用 Gibbs 抽样方法模拟生成 (X,Y) 的样本链。 □

例 3.7.6. 在 Gibbs 抽样中，每次变化的可以不是单个的分量，而是两个或多个分量。例如，设某个试验有 r 种不同结果，相应概率为 $\boldsymbol{p}=(p_1,\ldots,p_r)$(其中 $\sum_{i=1}^r p_i=1$)，独立重复试验 n 次，各个结果出现的次数 $\boldsymbol{X}=(X_1,\ldots,X_r)$ 服从多项分布。设 $A=\{X_1\ge 1,\ldots,X_r\ge 1\}$，假设 $P(A)$ 概率很小，要在条件 A 下对 $\boldsymbol{X}$ 抽样，如果先生成 $\boldsymbol{X}$ 的无条件样本再舍弃不符合条件 A 的部分则效率太低，可以采用如下的 Gibbs 抽样方法。

首先，任取初值 $\boldsymbol{X}^{(0)}$, 如 $\boldsymbol{X}^{(0)}=(1,\ldots,1)$。假设已生成了 $\boldsymbol{X}^{(t)}$，下一步首先从 $(1,\ldots,r)$ 中随机地抽取两个下标 (i,j), 令 $s=X_i^{(t)}+X_j^{(t)}$，在给定 $X_k, k\ne i,j$ 为 $\boldsymbol{X}^{(t)}$ 对应元素的条件下，(X_i,X_j) 的条件分布实际是 (X_i,X_j) 在 $X_i+X_j=s$ 以及 $X_i\ge 1, X_j\ge 1$ 条件下的分布。于是，在以上条件下，X_i 服从 $\mathrm{B}(s,p_i/(p_i+p_j))$ 分布限制在 $1\le X_i\le s-1$ 条件下的分布，即

$$\begin{aligned} q_k \triangleq & P(X_i=k|X_i+X_j=s, X_i\ge 1, X_j\ge 1) \\ =& \frac{C_s^k\left(\frac{p_i}{p_i+p_j}\right)^k\left(\frac{p_j}{p_i+p_j}\right)^{s-k}}{1-\left(\frac{p_j}{p_i+p_j}\right)^s-\left(\frac{p_i}{p_i+p_j}\right)^s},\ k=1,2,\ldots,s-1. \end{aligned}$$

要生成这样的 X_i 的抽样只要用生成离散型随机数的逆变换法。设抽取的 X_i 值为 X_i^*, 取 $(X_i^{(t+1)}, X_j^{(t+1)})=(X_i^*, s-X_i^*)$, 取 $\boldsymbol{X}^{(t+1)}$ 的其它元素为 $\boldsymbol{X}^{(t)}$ 的对应元素。如此重复就可以生成所需的 $\boldsymbol{X}$ 在条件 A 下的抽样链。 □

3.7.4 MCMC 计算软件 *

MCMC 是贝叶斯统计计算中最常用的计算工具。OpenBUGS 是一个成熟的 MCMC 计算开源软件（见 Lunn et al(2009)[29], Cowles(2013)[16], 另一个类似的有关软件是 WinBUGS[30]），能够进行十分复杂的贝叶斯模型的计算, 可以在 R 中直接调用 OpenBUGS 进行计算。

OpenBUGS 采用 Gibbs 抽样方法从贝叶斯后验分布中抽样，用户只需要指定先验分布和似然函数以及观测数据、已知参数，以及并行地生成多少个马氏链、链的一些初值、运行步数。软件自动计算 Gibbs 抽样所需的条件分布, 产生马氏链，并可以用图形和数值辅助判断收敛性，给出后验推断的概括统计。

在 R 中通过 BRugs 包调用 OpenBUGS 的功能。BRugs 用三类输入文件指定一个贝叶斯模型，第一类文件指定似然函数和参数先验密度，第二类文件指定已知参数、样本值，第三类文件指定马氏链初值（并行产生多个链时需要指定多组初值）。R 的 coda 软件包可以帮助对 MCMC 抽样结果进行分析和诊断。下面用一个简单例子介绍在 R 中用 BRugs 和 OpenBUGS 从贝叶斯后验中抽样的基本步骤。

例 3.7.7. 对例3.7.2，用 R 的 BRugs 包调用 OpenBUGS 来计算。设 $(X_1,\ldots,X_5)$ 的观测值为 $(74,85,69,17,5)$。

OpenBUGS 用模型文件描述随机变量分布和对参数的依赖关系，以及参数之间的关系。首先建立如下模型文件, 保存在文件 `pfl-model.txt` 中:

```
model
{
    p[1] <- 1/3
    p[2] <- (1-b)/3
    p[3] <- (1-2*b)/3
    p[4] <- 2*b/3
    p[5] <- b/3
    b <- b2 / 2
    b2 ~ dbeta(1,1)
    x[1:5] ~ dmulti(p[1:5], N)
}
```

文件中用向左的箭头表示确定性的关系，用方括号加序号表示下标，用 $\sim$ 表示左边的变量服从右边的分布。这里，b 为参数 β, b2 $= 2b$ 服从 Beta(1,1) 分布即 U(0,1) 分布，于是 β 有先验分布 U(0,0.5)。`x[1:5]` 表示向量 $(x_1,\ldots,x_5)$，服从多项分布，参数为 $(p_1,\ldots,p_5)$, 试验次数为 N。N 在模型文件中没有指定，将在数据文件中给出。

建立如下的数据文件，保存在文件 `pfl-data.txt` 中:

```
list(x=c(74, 85, 69, 17, 5),
    N=250)
```

这样的数据文件是 R 软件的列表格式，列表中的标量和向量为通常的 R 程序写法，矩阵用如

```
list(M=structure(.Data=c(1,2,3,4,5,6), .Dim=c(3,2)))
```

表示，其中.Dim 给出矩阵的行、列数（3×2），.Data 给出按行排列的所有元素（第一行为 1,2, 第二行为 3,4，第三行为 5,6）。

OpenBUGS 需要用户指定各个链的要抽样的参数的初值，这里我们要抽样的参数是 b, 但 b 是由 b2 计算得到的，所以对 b2 设置初值。设有如下两个初值文件，不同链的初值应尽可能不同，文件 pfl-inits1.txt 内容为:

```
list(b2=0.2)
```

文件 pfl-inits2.txt 内容为:

```
list(b2=0.8)
```

在 R 中，首先调用 BRugs 软件包:

```
require(BRugs)
```

然后，读入并检查模型文件：

```
modelCheck('pfl-model.txt')
```

读入数据文件:

```
modelData('pfl-data.txt')
```

下面，对模型和数据进行编译，得到抽样方案, 下面的语句要求并行运行两个链:

```
modelCompile(numChains=2)
```

准备迭代地生成 MCMC 抽样了，首先指定初值:

```
modelInits(c('pfl-inits1.txt', 'pfl-inits2.txt'))
```

下面，先试验性地运行 1000 次，作为老化期:

```
modelUpdate(1000)
```

现在才指定抽样要输出那些随机变量的随机数:

```
samplesSet(c('b', 'p[1:5]'))
```

现在可以抽样了，指定运行 10000 次，两个链并行运行:

```
modelUpdate(10000)
```

得到抽样后，可以把抽样的结果保存在 R 的变量中，比如

```
b2chains <- samplesHistory('b', plot=FALSE)
```

得到一个列表，有唯一的元素 b, 为 2×10000 的矩阵，每行是一个链的记录。sampleHistory 可以抽样链的曲线图，如

```
samplesHistory('b')
```

OpenBUGS 提供了一系列的简单统计和收敛诊断功能。如下程序列出各抽样变量的简单统计:

```
print(samplesStats("*"))
```

结果为:

```
        mean       sd  MC_error val2.5pc  median val97.5pc start sample
b    0.08757 0.016900 1.168e-04  0.05737 0.08668   0.12330  1001  20000
p[2] 0.30410 0.005632 3.892e-05  0.29230 0.30440   0.31420  1001  20000
p[3] 0.27500 0.011260 7.784e-05  0.25120 0.27550   0.29510  1001  20000
p[4] 0.05838 0.011260 7.784e-05  0.03824 0.05778   0.08217  1001  20000
p[5] 0.02919 0.005632 3.892e-05  0.01912 0.02889   0.04109  1001  20000
```

统计使用所有链的数据。可以看出，β 的后验均值为 0.08757。表中的 MC_error 表示估计后验均值时由于随机模拟导致的误差的标准差的估计，这个标准差估计针对抽样自相关性进行了校正。在老化期之后的运行次数越多 MC_error 越小，一个常用的经验规则是保证 MC_error 小于后验标准差的 5%，这里的结果提示还需要更多的运行次数。val2.5pc 和 val97.5pc

是抽样的后验分布的 2.5% 和 97.5% 分位数的估计值, 由此得到 β 的水平 95% 的可信区间 (credible interval) 为 $(0.05734, 0.1233)$。

如下程序画出 `b` 的后验密度估计:

```
samplesDensity('b', mfrow=c(1,1))
```

如下程序画抽样的 `b` 的自相关函数估计:

```
samplesAutoC('b', 1, mfrow=c(1,1))
```

当自相关函数很大而且衰减缓慢时生成的抽样链的效率较低。本例中的自相关函数基本表现为不相关列。

为了检查链是否收敛，OpenBUGS 提供了 BGR 统计量图:

```
samplesBgr('b', mfrow=c(1,1))
```

BGR 统计量的原理是考虑并行运行的每个链内部的变化情况，以及把所有的链合并在一起的变化情况，对这两种变化情况进行比较，当链收敛时，每个链内部的变化情况应该与合并在一起的变化情况很类似。类似于单因素方差分析中组间平方和与组内平方和的比较。当 BGR 图中的红色线接近于 1 并且三条线都保持稳定时就提示链收敛了。

□

3.8 序贯重要抽样 *

MCMC 是目前广泛使用的随机模拟方法，其中 Gibbs 抽样方法（见 §3.7.3）在确定了各个分量的条件分布后可以轮流产生各个分量的抽样。序贯重要抽样方法是重要抽样法（见 §3.2.4）的一种推广，其做法与 Gibbs 抽样方法有些相似，也是每次从一个分量抽样。

回顾多维随机变量抽样的条件分布法（见 §2.3.1）。设随机向量 $\boldsymbol{X}$ 的密度 $\pi(\boldsymbol{x}) = \pi(x_1, x_2, \ldots, x_n)$ 可以逐次地分解为条件密度乘积

$$\pi(x_1, x_2, \ldots, x_n) = \pi(x_1)\pi(x_2|x_1)\pi(x_3|x_1, x_2)\cdots\pi(x_n|x_1, \ldots, x_{n-1})$$

则可以用条件分布法产生 $\boldsymbol{X}$ 的抽样。当数据是时间序列或者可以依次增加对 π 的信息 (比如 π 是 Bayes 后验分布) 时这种方法很自然。但是，条件分布 $\pi(x_t|x_1, \ldots, x_{t-1})$ 可能是难以得到的或难以抽样的。为此，采用重要抽样思想：取一系列辅助分布 $\pi_t(\boldsymbol{x}_t) = \pi_t(x_1, \ldots, x_t), t =$

$1,\ldots,n$，使其近似于 $(X_1,\ldots,X_t)$ 的分布，$\pi_n=\pi$，各 π_t 可以分别有一个未知的归一化常数。对 $t=1,2,\ldots,n$ 用重要抽样法逐次抽取 $X_1,X_2,\ldots,X_n$，使得 $(X_1,\ldots,X_t)$ 是关于 π_t 适当加权的样本。抽取 X_t 时一般是给定 $X_1,\ldots,X_{t-1}$ 后从一个试抽样分布 $g_t(x_t|\boldsymbol{x}_{t-1})$ 抽取。适当计算权重就可以得到关于 π 适当加权的抽样 $\boldsymbol{X}_n=(X_1,X_2,\ldots,X_n)$。

这种抽样方法叫做**序贯重要抽样**（sequential importance sampling，SIS），算法如下:

置 $t\leftarrow 1$, 从 $g_1(\cdot)$ 抽取 X_1, 置 $W_1\leftarrow \pi_1(X_1)/g_1(X_1)$。

for(t **in** $2:n$) {

从 $g_t(x_t|X_1,\ldots,X_{t-1})$ 抽取 X_t, 记 $\boldsymbol{X}_t=(X_1,\ldots,X_{t-1},X_t)$;

计算**步进权重**

$$U_t\leftarrow \frac{\pi_t(\boldsymbol{X}_t)}{\pi_{t-1}(\boldsymbol{X}_{t-1})g_t(X_t|\boldsymbol{X}_{t-1})}$$

令 $W_t\leftarrow W_{t-1}U_t$;

}

输出 $(\boldsymbol{X}_n,W_n)$ 为关于 $\pi(\boldsymbol{x})$ 适当加权的样本。

SIS 一般同时独立进行 N 组，得到 $\{(\boldsymbol{X}_n^{(j)},W_n^{(j)})\}_{j=1}^N$，每一组称为一个“**流**”或一个“**粒子**”。

按照 SIS 步骤，有

$$\begin{aligned}
X_1\sim g_1(x_1), \quad & w_1=\frac{\pi_1(X_1)}{g_1(X_1)}\\
X_2\sim g_2(x_2|x_1), \quad & U_2=\frac{\pi_2(\boldsymbol{X}_2)}{\pi_1(X_1)g_2(X_2|X_1)},\\
& W_2=W_1U_2=\frac{\pi_2(\boldsymbol{X}_2)}{g_1(X_1)g_2(X_2|X_1)}\\
X_3\sim g_3(X_3|\boldsymbol{X}_2), \quad & U_3=\frac{\pi_3(\boldsymbol{X}_3)}{\pi_2(\boldsymbol{X}_2)g_3(X_3|\boldsymbol{X}_2)},\\
& W_3=\frac{\pi_3(\boldsymbol{X}_3)}{g_1(X_1)g_2(X_2|X_1)g_3(X_3|\boldsymbol{X}_2)}\\
& \cdots\cdots\\
X_n\sim g_n(x_n|\boldsymbol{X}_{n-1}), \quad & W_n=\frac{\pi_n(\boldsymbol{X}_n)}{g_1(X_1)g_2(X_2|X_1)\cdots g_n(X_n|\boldsymbol{X}_{n-1})}.
\end{aligned}$$

记

$$g_t(\boldsymbol{x}_t)=g_1(x_1)g_2(x_2|x_1)\cdots g_t(x_t|\boldsymbol{x}_{t-1})$$

则 $\boldsymbol{X}_t\sim g_t(\cdot)$ 而

$$W_t=\frac{\pi_t(\boldsymbol{X}_t)}{g_t(\boldsymbol{X}_t)},$$

因此 $(\boldsymbol{X}_t, W_t)$ 关于 π_t 适当加权 $(t = 1, \ldots, n)$。注意 $\pi_n = \pi$ 故这样得到的 $(\boldsymbol{X}_n, W_n)$ 关于 $\pi(\cdot)$ 适当加权。

试抽样分布的一种常见取法为 $g_t(x_t|\boldsymbol{x}_{t-1}) = \pi_t(x_t|\boldsymbol{x}_{t-1})$。

3.8.1 非线性滤波平滑

SIS 方法可以用在很多统计模型的计算中, 作为示例，考虑如下的非线性滤波平滑问题。

设不可观测的“状态”X_t 服从如下状态方程

$$X_t \sim q_t(\cdot|X_{t-1}, \theta),\ t = 1, 2, \ldots, n$$

X_t 的信息可以反映在可观测的 Y_t 中，其关系满足如下观测方程

$$Y_t \sim f_t(\cdot|X_t, \phi),\ t = 1, 2, \ldots, n$$

已知 $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n) = \boldsymbol{y}_n$ 和 θ, ϕ 后估计 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 的问题称为滤波平滑问题, 只需要求后验分布 $\pi(\boldsymbol{x}_n) = p_{\boldsymbol{X}_n|\boldsymbol{Y}_n}(\boldsymbol{x}_n|\boldsymbol{y}_n)$。如果能从 π 大量抽样 $\boldsymbol{X}^{(j)}, j = 1, \ldots, N$ 则可以用随机模拟方法对 $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ 的后验分布进行推断。下面用 SIS 方法产生 π 的样本。

记

$$\pi_t(\boldsymbol{x}_t) = p_{\boldsymbol{X}_t|\boldsymbol{Y}_t}(\boldsymbol{x}_t|\boldsymbol{y}_t),\ t = 1, 2, \ldots, n$$

注意

$$\pi_t(\boldsymbol{x}_t) \propto f_t(y_t|x_t) q_t(x_t|x_{t-1}) \pi_{t-1}(\boldsymbol{x}_{t-1}) \tag{3.98}$$

取试抽样分布 $g_t(x_t|\boldsymbol{x}_{t-1}) = q_t(x_t|x_{t-1})$, 对 $t = 1$, $\pi_1(x_1) = p_{X_1|Y_1}(x_1|y_1) \propto f_1(y_1|x_1)p(x_1)$, 其中 $p(x_1)$ 为 X_1 的分布密度，抽取 $X_1 \sim g_1(x_1)$, 如果可能应取 $g_1(x_1) = \pi_1(x_1)$。

产生关于 π 适当加权的 $\boldsymbol{X}_n$ 抽样的 SIS 步骤如下:

置 $t \leftarrow 1$, 从 $g_1(x_1)$ 抽取 X_1, 置 $W_1 \leftarrow \pi_1(X_1)/g_1(X_1)$。
for(t **in** $2 : n$) {
  从 $q_t(x_t|X_{t-1})$ 抽取 X_t, 记 $\boldsymbol{X}_t = (X_1, \ldots, X_{t-1}, X_t)$;
  令步进权重 $U_t \leftarrow f_t(y_t|X_t)$
  令 $W_t \leftarrow W_{t-1}U_t$;
}
输出 $(\boldsymbol{X}_n, W_n)$ 为关于 $\pi(\boldsymbol{x})$ 适当加权的样本。

这种方法相当于用状态方程前进一步获得下一分量的抽样，用同时刻的观测值 y_t 的似然作为步进权重。这样，t 步以后得到的 $\boldsymbol{X}_t$ 服从

$$g_t(\boldsymbol{x}_t) = g_{t-1}(\boldsymbol{x}_{t-1})q_t(x_t|x_{t-1})$$

其中 $g_1(x_1) \propto f_1(y_1|x_1)$。于是

$$\begin{aligned}\frac{\pi_t(\boldsymbol{x}_t)}{g_t(\boldsymbol{x}_t)} &= \frac{f_t(y_t|x_t)q_t(x_t|x_{t-1})\pi_{t-1}(\boldsymbol{x}_{t-1})}{g_{t-1}(\boldsymbol{x}_{t-1})q_t(x_t|x_{t-1})} \\ &= f_t(y_t|x_t)\frac{\pi_{t-1}(\boldsymbol{x}_{t-1})}{g_{t-1}(\boldsymbol{x}_{t-1})}, \\ W_t &= f_t(y_t|X_t)W_{t-1} = \frac{\pi_t(\boldsymbol{X}_t)}{g_t(\boldsymbol{X}_t)},\end{aligned}$$

可见 $(\boldsymbol{X}_t, W_t)$ 关于 π_t 适当加权, $(\boldsymbol{X}_n, W_n)$ 关于 $\pi_n(\boldsymbol{x}) = p_{\boldsymbol{X}_n|\boldsymbol{Y}_n}(\boldsymbol{x}_n|\boldsymbol{y}_n)$ 适当加权。

设第 t 步的抽样为 $\boldsymbol{X}_t^{(i)}, i = 1, \ldots, N$, 对应权重为 $W_t^{(i)}, i = 1, \ldots, n$。以上的方法在抽取 $(X_1, \ldots, X_n)$ 时不考虑 $(y_1, \ldots, y_n)$ 的具体取值，这样的试抽样分布虽然很容易抽样，但是效果很差，权重 $\{W_t^i, i = 1, \ldots, N\}$ 随着 t 的增加会把变得差异很大，以至于 N 个流中只有极少数流能起作用。

由(3.98)可见

$$\pi_t(x_t|\boldsymbol{x}_{t-1}) = p_{X_t|\boldsymbol{X}_{t-1},Y_t}(x_t|\boldsymbol{x}_{t-1}) \propto f_t(y_t|x_t)q_t(x_t|x_{t-1}), \tag{3.99}$$

如果能取试抽样分布 $g_t(x_t) \propto f_t(y_t|x_t)q_t(x_t|x_{t-1})$ 则每次抽取 X_t 都利用了同期的观测值 y_t 的信息，会大大改善抽样效率。更进一步，设 $\pi_{t+1}(\boldsymbol{x}_{t+1})$ 关于 $\boldsymbol{x}_t$ 的边缘分布为 $\pi_{t,t+1}(\boldsymbol{x}_t)$, 如果在第 t 步能从 $\pi_{t,t+1}(\boldsymbol{x}_t)$ 的条件分布 $p(x_t|\boldsymbol{x}_{t-1})$ 抽样，就可以利用 y_t, y_{t+1} 的信息，抽样效率可以进一步改善。

另外一种改进的办法是再抽样，增加权重大的流，舍弃权重小的流。

3.8.2 再抽样

如果试抽样分布选取不适当，最后的权重可能会差别很大，体现在权重 $\{W_t^{(i)}, i = 1, \ldots, N\}$ 的样本变异系数很大，称为权重偏斜严重。这时，权重小的流基本不起作用。出现这样的情况时，可以把一些权重太小的流舍弃而增加权重大的流，这样的技术称为**再抽样**。

简单随机再抽样

设进行 SIS 时在时刻 t 已经得到了 N 个部分的流 $\{\boldsymbol{X}_t^{(i)}, i = 1, \ldots, N\}$ 以及相应的权重 $\{W_t^{(i)}, i = 1, \ldots, N\}$。可以在每一步都按照权重对流再抽样，也可以仅当权重偏斜严重

时才进行再抽样。如果在第 t 步再抽样，只要以正比于 $W_t^{(i)}$ 的概率从 $\{\boldsymbol{X}_t^{(j)}, j=1,\ldots,N\}$ 中抽取 $\boldsymbol{X}_t^{(i)}$, 独立有放回地抽取 N 个，记作 $\{\boldsymbol{X}_t^{*(i)}, i=1,\ldots,N\}$, 并把权重调整为相等的 $W_t^{*(i)}=\frac{1}{N}\sum_{j=1}^N W_t^{(j)}$, $i=1,\ldots,N$。这样的方法称为**简单随机再抽样**。这样再抽样后各个流不再是独立的。

剩余再抽样

为了达到以上的简单随机抽样的效果，还可以把大权重的的流直接复制多份，小权重的流仅以一定比例保留，其它舍弃。这种方法称为**剩余再抽样** (residual resampling)，其计算量更小而且模拟误差更小。算法描述如下。如果在第 t 步需要再抽样，则计算 $\bar{W}_t=\frac{1}{N}\sum_{j=1}^N W_t^{(j)}$, 然后按如下做法从 N 个流中重新抽取 N 个。首先，对 $i=1,\ldots,N$, 直接保留 $k_i=[W_t^{(i)}/\bar{W}_t]$ 份 $\boldsymbol{X}_t^{(i)}$([·] 表示向下取整)；其次，令 $N_r=N-\sum_{i=1}^N k_i$ 为缺额个数，随机有放回地按照正比于 $\frac{W_t^{(i)}}{\bar{W}_t}-k_i$(取整后的小数部分) 的概率从 $\{\boldsymbol{X}_t^{(j)}, j=1,\ldots,N\}$ 中抽取 $\boldsymbol{X}_t^{(i)}$, 共抽取 N_r 个。这样得到了 N 个新的流, 记作 $\{\boldsymbol{X}_t^{*(i)}, i=1,\ldots,N\}$, 并调整其权重为相等的 $W_t^{*(i)}=\bar{W}_t$, $i=1,\ldots,N$。这种方法的第一步把权重大的流直接复制了若干份，第二步对按剩余的权重再抽取到满 N 个流为止。

舍选控制再抽样

简单随机再抽样和剩余再抽样都使得结果各个流不再独立。另外一种想法是采用 §3.2.4中介绍的舍选控制方法，对权重小的流从头重新抽样并适当调整权重。首先，设定若干个要执行再抽样的时间点 $0<t_1<\cdots<t_k\le n$, 以及相应的权重阈值 $c_1,\ldots,c_k$。在 $t=t_j$ 时，进行舍选控制再抽样。若流 i 的权重 $W_t^{(i)}\ge c_j$, 则保留此流 $\boldsymbol{X}_t^{(i)}$ 和权重 $W_t^{(i)}$ 不变；若流 i 的权重 $W_t^{(i)}<c_j$, 则以概率 $W_t^{(i)}/c_j$ 保留此流，如果决定保留，则修改其权重为 c_j，如果决定舍弃此流，则从 $t=1$ 重新生成这个流，同样也需要经过 $t_1,\ldots,t_j$ 处的舍选判断，如果被舍弃就还从 $t=1$ 重新开始，直到被接受。

部分舍选控制再抽样

如果从头开始重新抽样的情况发生比较多则模拟的效率会比较低，为此可以采用如下的部分舍选控制再抽样的 SIS 方法。

首先，设定若干个要执行再抽样的时间点 $0<t_1<\cdots<t_k\le n$, 以及相应的权重阈值 $c_1,\ldots,c_k$。在 $t=t_j$ 时，进行部分舍选控制再抽样。若流 i 的权重 $W_t^{(i)}\ge c_j$, 则保留此流 $\boldsymbol{X}_t^{(i)}$ 和权重 $W_t^{(i)}$ 不变。若流 i 的权重 $W_t^{(i)}<c_j$, 则以概率 $W_t^{(i)}/c_j$ 保留此流, 如果决定保留，则修改其权重为 c_j。如果决定舍弃此流，则不是从头重新生成这个流，而是按照概率

正比于 $W_{t_{j-1}}^{(s)}$ 从 $\{\boldsymbol{X}_{t_{j-1}}^{(s)}, s=1,\ldots,N\}$ 中随机抽取一个替换原来的 $\boldsymbol{X}_{t_{j-1}}^{(i)}$, 用 t_{j-1} 时的权重 $\{W_{t_{j-1}}^{(s)}\}$ 的平均值 $\bar{W}_{t_{j-1}}$ 代替原来的权重 $W_{t_{j-1}}^{(i)}$，然后继续按照 SIS 标准步骤将此流经 $t=t_{j-1}+1,\ldots,t_j$ 延伸得到新的 $\boldsymbol{X}_{t_j}^{(i)}$ 和权重 $W_{t_j}^{(i)}$，然后再进行 t_j 处的舍选控制，如果被舍弃就再从 t_{j-1} 处随机再抽样后继续，直到在 t_j 处被接受。

关于序贯重要抽样更详细的讨论参见 Liu(2001)[28]。

习题三

1. 设 $\sigma_j, j=1,2,\ldots,m$ 为 m 个正实数,

$$f(\alpha_1,\alpha_2,\ldots,\alpha_m)=\frac{\sigma_1^2}{\alpha_1}+\frac{\sigma_2^2}{\alpha_2}+\cdots+\frac{\sigma_m^2}{\alpha_m},\ (\alpha_1,\ldots,\alpha_m)\in(0,1)^m,$$

则在 $\alpha_1+\alpha_2+\cdots+\alpha_m=1$ 条件下 $f(\alpha_1,\alpha_2,\ldots,\alpha_m)$ 的最小值点为

$$\alpha_j=\frac{\sigma_j}{\sum_{k=1}^m\sigma_k},\ j=1,2,\ldots,m$$

最小值为 $(\sigma_1+\cdots+\sigma_m)^2$。

2. 考虑定积分

$$I=\int_{-1}^{1}e^x dx=e-e^{-1}.$$

(1) 用随机模拟方法计算定积分 I, 分别用随机投点法、平均值法、重要抽样法和分层抽样法计算。

(2) 设估计结果为 $\hat{I}$, 如果需要以 95% 置信度保证计算结果精度在小数点后三位小数，这四种方法分别需要计算多少次被积函数值？

(3) 用不同的随机数种子重复以上的估计 B 次，得到 $\hat{I}_j, j=1,2,\ldots,B$, 由此估计 $\hat{I}$ 的抽样分布方差，与 (2) 的结果进行验证。

(4) 称

$$\mathrm{MAE}(\hat{I})=E|\hat{I}-I|$$

为 $\hat{I}$ 的平均绝对误差。从 (3) 得到的 $\hat{I}_j, j=1,2,\ldots,B$ 中估计 $\mathrm{MAE}(\hat{I})$。比较这四种积分方法的平均绝对误差大小。

3. 设 $h(x)=\frac{e^{-x}}{1+x^2}$, $x\in(0,1)$, 用重要抽样法计算积分 $I=\int_0^1 h(x)\,dx$，分别采用如下的试抽样密度:

$$\begin{aligned}
f_1(x) =& 1,\ x\in(0,1),\\
f_2(x) =& e^{-x},\ x\in(0,\infty),\\
f_3(x) =& \frac{1}{\pi(1+x^2)},\ x\in(-\infty,\infty),\\
f_4(x) =& (1-e^{-1})^{-1}e^{-x},\ x\in(0,1),\\
f_5(x) =& \frac{4}{\pi(1+x^2)},\ x\in(0,1).
\end{aligned}$$

   (1) 作 $h(x)$ 和各试抽样密度的图形，比较其形状。

   (2) 取样本点个数 $N=10000$, 分别给出对应于不同试抽样密度的估计 $\hat{I}_k$, $k=1,2,3,4,5$, 以及 $\mathrm{Var}(\hat{I}_k)$ 的估计。

   (3) 分析 $\mathrm{Var}(\hat{I}_k)$ 的大小差别的原因。

   (4) 把 $(0,1)$ 区间均分为 10 段，在每一段内取 $N=1000$ 个样本点用平均值法计算积分值，把各段的估计求和得到 I 的估计 $\hat{I}_6$，估计其方差。

   (5) 用例3.2.3的分层抽样方法计算积分的估计 $\hat{I}_7$，估计 $\mathrm{Var}(\hat{I}_7)$ 并与前面的结果进行比较。

4. 设 $X\sim \mathrm{N}(0,1)$, $h(x)=\exp(-\frac{1}{2}(x-3)^2)+\exp(-\frac{1}{2}(x-6)^2)$。令 $I=Eh(X)$。

   (1) 推导 I 的精确表达式并计算结果。

   (2) 用 $N=1000$ 次函数计算的平均值法估计 I 并估计误差大小。

   (3) 设计适当的重要抽样方法取 $N=1000$ 估计 I 并估计误差大小。

5. 设 $X\sim \mathrm{N}(0,1)$, 则 $\theta=P(X>4.5)=3.398\times 10^{-6}$。

   (1) 如果直接生成 N 个 X 的随机数，用 $X_i>4.5$ 的比例估计 $P(X>4.5)$，平均多少个样本点中才能有一个样本点满足 $X_i>4.5$?

   (3) 取 V 为指数分布 $\mathrm{Exp}(1)$, 令 $W=V+4.5$，用 W 的样本进行重要抽样估计 θ, 取样本点个数 $N=1000$，求估计值并估计误差大小。

6. 设 $\{U_i, i=1,2,\dots\}$ 为独立同 U(0,1) 分布的随机变量序列。令

$$K=\min\left\{k:\sum_{i=1}^{k}U_i>1\right\}.$$

(1) 证明 $EK = e$。

(2) 生成 K 的 N 个独立抽样，用平均值 $\bar{K}$ 估计 e。

(3) 估计 $\bar{K}$ 的标准差，给出 e 的近似 95% 置信区间。

7. 设 $\{U_i, i = 1, 2, \dots\}$ 为独立同 U(0,1) 分布的随机变量序列。令 M 为序列中第一个比前一个值小的元素的序号，即

$$M = \min\{m : U_1 \le U_2 \le \cdots \le U_{m-1}, U_{m-1} > U_m,\ m \ge 2\}$$

(1) 证明 $P(M > n) = \frac{1}{n!}$, $n \ge 2$。

(2) 用概率论中的恒等式 $EM = \sum_{n=0}^{\infty} P(M > n)$ 证明 $EM = e$。

(3) 生成 M 的 N 个独立抽样，用平均值 $\bar{M}$ 估计 e。

(4) 估计 $\bar{M}$ 的标准差，给出 e 的近似 95% 置信区间。

8. 在例3.2.5中，设 $n = 10$, 已知 (n_j, y_j) 的值为:

n_j	20	30	25	30	40	20	50	30	20	20
y_j	6	1	1	0	5	4	1	8	4	7

编写 R 程序估计 $E(\log K | y_1, y_2, \dots, y_n)$。

9. 用随机模拟法计算二重积分 $\int_0^1 \int_0^1 e^{(x+y)^2}\, dy\, dx$, 用对立变量法改善精度。

10. 设 $W \sim \chi^2(n-1)$，求 $\rho\left[(\frac{1}{n-1}W - 1)^2, (\frac{1}{n}W - 1)^2\right]$。

11. 用 R 程序实现例3.3.8的模拟。取 $\mu = 0$, $\sigma = 1$, $N = 100000$, $n = 5, 10, 30$，列出各偏差、均方误差和 $\hat{s}_1 - \hat{s}_2$ 的值。

12. 选择适当的编程语言实现例3.4.2的算法。考虑如下的变化情形:

(1) 银行 8:00 开门，17:00 关门，关门后不再允许顾客进入但已进入的顾客会服务完;

(2) 顾客到来服从非齐次的泊松分布，其速率函数 $\lambda(t)$ 为阶梯函数;

(3) 如果顾客到来后发现队太长，有一定概率离开；如果顾客等待了太长时间，有一定概率离开，这时需要求这些离开顾客的平均人数。

13. 设计模拟如下离散事件的算法。设某商场每天开放 L_0 时间，有扒手在该商城出没，扒手的出现服从强度 λ_1 的齐次泊松过程，出现后作案 X 时间后离开，设 X 服从对数正态分布, $\ln X \sim \mathrm{N}(\mu_2, \sigma_2^2)$。设有一个警察每隔 L_3 时间在商城巡逻 Y 时间，Y 服从 $\mathrm{Exp}(\lambda_2)$ 分布，只要扒手和警察同时出现在商城内扒手就会被抓获。模拟估计一天时间内有扒手被抓获的概率。

14. 设计模拟 M/M/c 随机服务系统的算法。设服务机构共有 c 个服务窗口，顾客到来服从齐次泊松过程，速率为 λ，顾客排成一队，一旦某个窗口空闲则队头的顾客接受服务，每个窗口的服务时间均服从期望为 $1/\mu$ 的指数分布。

    (1) 用 R 程序编程模拟上述随机服务系统；

    (2) 模拟估计顾客在这个服务机构的平均滞留时间 ER;

    (3) 在什么条件下这个服务机构的平均滞留时间能够稳定不变？

    (4) R 程序在处理大量循环方面比较慢。设法对程序效率进行改善。

15. §3.5给出了用模拟方法比较置信区间性能的步骤。设 $X \sim \mathrm{b}(1,p)$, $X_1, X_2, \ldots, X_n$ 为样本。令 $S_0 = \sum_{i=1}^n X_i$, $\hat{p} = S_0/n = \frac{1}{n}\sum_{i=1}^n X_i$。用模拟方法比较如下五种置信区间：

    (1) 利用正态近似。当 n 很大时

    $$\frac{\hat{p}-p}{\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}}$$

    近似服从 $\mathrm{N}(0,1)$，于是得置信区间

    $$\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}.$$

    (2) 利用正态近似，令

    $$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\hat{p})^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}),$$

    n 很大时

    $$\frac{\hat{p}-p}{\sqrt{\frac{1}{n}S^2}}$$

    近似服从 $\mathrm{N}(0,1)$，于是得置信区间

    $$\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}S^2}.$$

(3) Wilson 置信区间。利用正态近似，n 很大时

$$\frac{\hat{p}-p}{\sqrt{\frac{1}{n}p(1-p)}}$$

近似服从 $\mathrm{N}(0,1)$，解关于 p 的不等式

$$\left|\frac{\hat{p}-p}{\sqrt{\frac{1}{n}p(1-p)}}\right| \le z_{1-\frac{\alpha}{2}},$$

得置信区间 ($\lambda = z_{1-\frac{\alpha}{2}}$)

$$\frac{\hat{p}+\frac{\lambda^2}{2n}}{1+\frac{\lambda^2}{n}} \pm \frac{\lambda\sqrt{\frac{\lambda^2}{4n}+\hat{p}(1-\hat{p})}}{\sqrt{n}\left(1+\frac{\lambda^2}{n}\right)}.$$

(4) Agresti and Coull(1998) 方法。先计算 Wilson 区间中心 ($\lambda = z_{1-\frac{\alpha}{2}}$)

$$\tilde{p} = \frac{\hat{p}+\frac{\lambda^2}{2n}}{1+\frac{\lambda^2}{n}},$$

取置信区间为

$$\tilde{p} \pm \lambda\frac{\sqrt{\tilde{p}(1-\tilde{p})}}{\sqrt{n}\sqrt{1+\frac{\lambda^2}{n}}}. \tag{3.100}$$

(5) 用二项分布导出 p 的置信区间。令

$$p_1 = \left\{1+\frac{n-S_0+1}{S_0}F_{1-\frac{\alpha}{2}}(2(n-S_0+1),2S_0)\right\}^{-1}$$
$$p_2 = \left\{1+\frac{n-S_0}{S_0+1}\frac{1}{F_{1-\frac{\alpha}{2}}(2(S_0+1),2(n-S_0))}\right\}^{-1}$$

取置信区间为 $[p_1,p_2]$，其中 $F_q(n_1,n_2)$ 表示 $\mathrm{F}(n_1,n_2)$ 分布的 q 分位数。

对不同 $n,p,1-\alpha$，比较这五种置信区间的优劣。

16. 不同的检验法有不同的功效，在难以得到功效函数的显式表达式的时候，模拟方法可以起到重要补充作用。对无截距项的回归模型

$$y_i = bx_i+\varepsilon_i, \quad i=1,2,\ldots,n$$
$$\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n \overset{\text{iid}}{\sim} \mathrm{N}(0,\sigma^2)$$

为检验 $H_0: b=0$，有如下两种检验方法：

(1) $b=0$ 时 $y_i=\varepsilon_i$, 于是在 H_0 下

$$t=\frac{\bar{y}}{\sqrt{\frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

服从 $\mathrm{t}(n-1)$ 分布。设 λ 为 $\mathrm{t}(n-1)$ 分布的 $1-\frac{\alpha}{2}$ 分位数，取否定域为 $\{|t|>\lambda\}$。

(2) 令

$$\hat{b}=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2}, \qquad U=\hat{b}^2\sum_{i=1}^{n}x_i^2$$

$$Q=\sum_{i=1}^{n}y_i^2-U, \qquad F=\frac{U}{Q/(n-1)}$$

设 λ' 为 $\mathrm{F}(1,n-1)$ 的 $1-\alpha$ 分位数，取否定域为 $\{F>\lambda'\}$。

对不同的 b,σ^2,n,α 以及不同的 $\{x_i\}$ 模拟比较这两种检验方法的功效。

17. 对例3.6.7, 证明 $\hat{\phi}_1$ 的均方误差为

$$L_1=E(\hat{\phi}_1-\mu^2)^2=\frac{4\sigma^2\mu^2}{n}+\frac{2n}{n-1}\frac{\sigma^4}{n^2}.$$

18. 考虑下的非线性回归模型 (logistic 曲线)

$$y=A(1+e^{-bx})+\varepsilon,\ \varepsilon\sim\mathrm{N}(0,\sigma^2),$$

其中 $A>0,b>0,\sigma^2>0$ 是未知参数。设有独立样本 $(x_i,y_i),i=1,2,\ldots,n$, 可以用最小二乘方法估计 A,b,σ^2, 估计值记为 $\hat{A},\hat{b},\hat{\sigma}^2$。

(1) 设真实的 $A=10,b=1,\sigma=1$。编写 R 程序模拟一组样本（取 $x_i=-10,-9,\ldots,9,10$），计算 $\hat{A},\hat{b},\hat{\sigma}^2$，然后用 bootstrap 方法估计 $\hat{A},\hat{b},\hat{\sigma}^2$ 的标准误差和偏差。

(2) 重复生成 N 组模型的样本，从这 N 组样本中分别得到估计值 $\hat{A}^{(j)},\hat{b}^{(j)},(\hat{\sigma}^{(j)})^2$，$j=1,2,\ldots,N$, 用得到的这些估计值作为 $\hat{A},\hat{b},\hat{\sigma}^2$ 的抽样分布的样本，估计其标准误差和偏差并与 bootstrap 方法得到的结果进行对比。

19. 设 $\{X_t,t=0,1,\ldots\}$ 为状态取值于 $S=\{1,2,\ldots,m\}$ 的马氏链，有平稳分布 $\pi_j,j=1,2,\ldots,m$，若 X_0 服从平稳分布，证明所有 X_t 都服从平稳分布 $(t\geq 0)$。

20. 在例3.7.1中，设 $n=100$, $a=10000$, 编写 R 程序产生 MCMC 抽样链。

21. 在例3.7.2中，设 $(x_1,x_2,x_3,x_4,x_5)=(82,72,45,34,17)$, 选取适当的预热期和模拟，编写 R 程序计算 β 的后验均值估计。从不同初值出发多做几次考察估计的误差大小。

22. 在例3.7.3中，设 $A=20,B=10,d=1,K=11$，选取适当的预热期和步幅 σ，编写 R 程序产生 MCMC 抽样链。

23. 设向量 $\boldsymbol{x}=(x_1,\ldots,x_n)$, 每个 x_i 取值为 $+1$ 或 -1，记这样的 $\boldsymbol{x}$ 的集合为 $\mathcal{X}$。设随机向量 $\boldsymbol{X}$ 有如下概率质量函数

$$\pi(\boldsymbol{x})=P(\boldsymbol{X}=\boldsymbol{x})=\frac{1}{Z}\exp\left\{\mu\sum_{i=1}^{n-1}x_ix_{i+1}\right\}$$

其中 μ 已知，Z 为归一化常数。这样的模型称为一维 Ising 模型。取 $\mu=1,n=50$, 设计 MCMC 算法从 $\pi(\boldsymbol{x})$ 抽样。

24. 证明例3.7.4的(3.96)式。

25. 对例3.7.5, 取 $n=20$, $\alpha=\beta=0.5$, 生成 (X,Y) 的 Gibbs 抽样链，比较 Y 的样本的直方图和 Beta(α,β) 分布密度。

26. 设随机变量 X 和 Y 都取值于 $(0,B)$ 区间 (B 已知)。设 $Y=y$ 条件下 X 的条件分布密度为

$$f(x|y)\propto e^{-yx},\ x\in(0,B),$$

$X=x$ 条件下 Y 的条件分布密度为

$$f(y|x)\propto e^{-xy},\ y\in(0,B),$$

编写 R 程序用 Gibbs 抽样方法对 (X,Y) 抽样，估计 EX 和 $\rho(X,Y)$。

27. 设 $\boldsymbol{X}=(X_1,X_2,\ldots,X_n)$ 独立同 U(0,1) 分布，对 $0<d<\frac{1}{n-1}$, 用条件 A 表示这 n 个点两两的距离都超过 d，可以证明 A 的概率为 $[1-(n-1)d]^n$。设 $n=9$, $d=0.1$，设计 Gibbs 抽样方法生成满足条件 A 的 $\boldsymbol{X}$ 的抽样链。

28. 在调相通讯中，考虑如下的状态空间模型

$$\begin{aligned}X_t=&\phi_1X_{t-1}+\eta_t,\ \eta_t\sim N(0,\sigma_\eta^2),\ t=1,2,\ldots,n\\ y_t=&A\cos(ft+X_t)+\varepsilon_t,\ \eta_t\sim N(0,\sigma_\varepsilon^2),\ t=1,2,\ldots,n\end{aligned}\qquad(*)$$

其中 $\phi_1 = 0.6, \sigma_\eta^2 = 1/6, A = 320, f = 1.072 \times 10^7, \sigma_\varepsilon^2 = 1$, $(y_1, \ldots, y_n)$ 为观测值，$(X_1, \ldots, X_n)$ 为不可观测的随机变量。

(1) 设 $X_0 = 0$，$n = 128$，模拟生成 $(X_t, y_t), t = 1, 2, \ldots, n$。

(2) 根据 §3.8.1 设计 SIS 算法产生关于已知 $y_1, \ldots, y_n$ 条件下 $X_1, \ldots, X_n$ 的条件分布的适当加权样本，共生成 $N = 10000$ 组, 试抽样采用从 (*) 向前一步的方法。

(3) 考察以上得到的权重 $\{W_i\}$ 的分布情况。

(4) 在 SIS 抽样的每一步进行剩余再抽样；

(5) 根据后验均值方法利用上述改进的抽样估计 $(X_1, \ldots, X_n)$;

(6) 对每个 X_t，计算上述后验估计的标准误差；

(7) 独立地重复 $M = 400$ 次估计过程，从 M 次不同的后验估计计算新的估计标准误差, 与 (6) 得到的结果进行比较。

第四章　近似计算

在统计计算和其它科学计算中，经常需要计算各种函数的值，对函数进行逼近，用数值方法计算积分、微分。本章讨论这些计算问题。

4.1　函数逼近 *

4.1.1　多项式逼近

数学中的超越函数如 e^x、$\ln x$、$\sin x$ 在计算机中经常用泰勒级数展开来计算，这就是用多项式来逼近函数。多项式的高效算法见例1.3.1。数学分析中的 Weirstrass 定理表明，闭区间上的连续函数可以用多项式一致逼近。泰勒展开要求函数有多阶导数，我们需要找到对更一般函数做多项式逼近的方法。

考虑如下的函数空间

$$L^2[a,b] = \{g(\cdot) : g(x)\text{定义于}[a,b], \text{且} \int_a^b g^2(x)w(x)dx < \infty\} \tag{4.1}$$

则 $L^2[a,b]$ 是线性空间，在 $L^2[a,b]$ 中定义内积

$$< f, g >= \int_a^b f(x)g(x)w(x)dx, \tag{4.2}$$

则 $L^2(a,b)$ 为 Hilbert 空间。对 $g(x) \in L^2[a,b]$, 假设希望用 n 阶多项式 $f_n(x)$ 逼近 $g(x)$，使得

$$\|f_n - g\|^2 = \int_a^b |f_n(x) - g(x)|^2 w(x)dx \tag{4.3}$$

最小。如何求这样的多项式？

用 Gram-Schmidt 正交化方法可以在 $L^2[a,b]$ 中把多项式序列 $\{1, x, x^2, \dots\}$ 正交化为正交序列 $\{P_0, P_1, P_2, \dots\}$，序列中函数彼此正交，且 P_k 是 k 阶多项式, 称 $\{P_0, P_1, P_2, \dots\}$ 为正交

多项式。设 $H_n[a,b]$ 为函数 $\{1,x,x^2,\ldots,x^n\}$ 的线性组合构成的线性空间，则 $\{P_0,P_1,\ldots,P_n\}$ 构成 $H_n[a,b]$ 的正交基且 $P_n[a,b]$ 是 $L^2[a,b]$ 的子 Hilbert 空间，使得加权平方距离(4.3)最小的 $f_n(x)$ 是 $g(\cdot)$ 在子空间 $H_n[a,b]$ 的投影，记为 $\mathrm{Proj}_{H_n[a,b]}(g)$, 投影可以表示为 $\{P_0,P_1,\ldots,P_n\}$ 的线性组合

$$\underset{H_n[a,b]}{\mathrm{Proj}}\,(g)=\sum_{j=0}^{n}\frac{<g,P_j>}{\|P_j\|^2}P_j. \tag{4.4}$$

这样，只要预先找到 $[a,b]$ 上的多项式的正交基，通过计算内积就可以很容易地找到使得(4.3)最小的 $f_n(x)$。对于 $L^2[a,b]$ 中的任意函数 $g(x)$ 有

$$\lim_{n\to\infty}\|\underset{H_n[a,b]}{\mathrm{Proj}}\,(g)-g\|^2=0, \tag{4.5}$$

于是有

$$g=\lim_{n\to\infty}\underset{H_n[a,b]}{\mathrm{Proj}}\,(g)=\sum_{j=0}^{\infty}\frac{<g,P_j>}{\|P_j\|^2}P_j. \tag{4.6}$$

因为 $L^2[a,b]$ 依赖于定义域 $[a,b]$ 和权重函数 $w(\cdot)$，所以正交多项式也依赖于 $[a,b]$ 和 $w(\cdot)$。针对定义域 $[-1,1]$, $[0,\infty)$ 和 $(-\infty,\infty)$ 和几种不同的权重函数可以得到不同的正交多项式序列，表4.1列出了这些正交多项式的定义域、权重函数和名称。

表 4.1: 不同定义域和权重函数的正交多项式

函数空间	权函数	正交多项式	记号
$L^2[-1,1]$	1	Legendre	$P_n(x)$
$L^2[-1,1]$	$(1-x^2)^{-1/2}$(边界加重)	Chebyshev I 型	$T_n(x)$
$L^2[-1,1]$	$(1-x^2)^{1/2}$(中心加重)	Chebyshev II 型	$U_n(x)$
$L^2[0,\infty)$	$\exp(-x)$	Laguerre	$L_n(x)$
$L^2(-\infty,\infty)$	$\exp(-x^2)$	Hermite	$H_n(x)$
$L^2(-\infty,\infty)$	$\exp(-x^2/2)$	修正 Hermite	$He_n(x)$

Legendre 多项式是权重函数 $w(x)=1$ 的 $L^2[-1,1]$ 中的正交多项式，此权重函数在整个区间是等权重的，内积为

$$<f,g>=\int_{-1}^{1}f(x)g(x)\,dx.$$

有一般公式

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n,$$

整个序列可以递推计算如下：

$$\begin{aligned} P_0(x) =& 1, \quad P_1(x) = x, \quad P_2(x) = (3x^2-1)/2, \\ P_3(x) =& (5x^3-3x)/2, \quad P_4(x) = (35x^4-30x^2+3)/8, \\ (n+1)P_{n+1}(x) =& (2n+1)xP_n(x) - nP_{n-1}(x), \end{aligned}$$

且 $\|P_n\|^2 = 2/(2n+1)$。

Chebyshev I 型多项式是权重函数为 $w(x) = (1-x^2)^{-1/2}$ 的 $L^2[-1,1]$ 中的正交多项式，此权重函数强调两端，内积为

$$< f, g >= \int_{-1}^{1} f(x)g(x)(1-x^2)^{-1/2}\,dx,$$

$T_n(x)$ 有一般表达式 $\cos(n\cos^{-1}(x))$, 整个序列可以递推计算如下：

$$\begin{aligned} T_0(x) =& 1, \quad T_1(x) = x, \quad T_2(x) = 2x^2-1, \\ T_3(x) =& 4x^3-3x, \quad T_4(x) = 8x^4-8x^2+1, \\ T_{n+1}(x) =& 2xT_n(x) - T_{n-1}(x) \end{aligned}$$

且 $\|T_0\|^2 = \pi$, $\|T_n\|^2 = \frac{\pi}{2}(n>0)$。

Chebyshev II 型多项式是权重函数为 $w(x) = (1-x^2)^{1/2}$ 的 $L^2[-1,1]$ 中的正交多项式，此权重函数强调区间中间，内积为

$$< f, g >= \int_{-1}^{1} f(x)g(x)(1-x^2)^{1/2}\,dx,$$

整个序列可以递推计算如下：

$$\begin{aligned} U_0(x) =& 1, \quad U_1(x) = 2x, \quad U_2(x) = 4x^2-1, \\ U_3(x) =& 8x^3-4x, \quad U_4(x) = 16x^4-12x^2+1, \\ U_{n+1}(x) =& 2xU_n(x) - U_{n-1}(x), \end{aligned}$$

且 $\|U_n\|^2 = \pi/2$。

Laguerre 多项式是权重函数为 $w(x)=e^{-x}$ 的 $L^2[0,\infty)$ 中的正交多项式，此权重函数强调区间左边，内积为

$$< f,g >= \int_0^{\infty} f(x)g(x)e^{-x}\,dx,$$

整个序列可以递推计算如下：

$$\begin{aligned} L_0(x) =& 1, \quad L_1(x)=1-x, \quad L_2(x)=(2-4x+x^2)/2,\\ L_3(x) =& (6-18x+9x^2-x^3)/6,\\ L_4(x) =& (24-96x+72x^2-16x^3+x^4)/24,\\ (n+1)L_{n+1}(x) =& (2n+1-x)L_n(x)-nL_{n-1}(x), \end{aligned}$$

且 $\|L_n\|^2=1$。

Hermite 多项式是权重函数为 $w(x)=e^{-x^2}$ 的 $L^2(-\infty,\infty)$ 中的正交多项式，此权重函数强调区间中央，内积为

$$< f,g >= \int_{-\infty}^{\infty} f(x)g(x)e^{-x^2}\,dx,$$

整个序列可以递推计算如下：

$$\begin{aligned} H_0(x) =& 1, \quad H_1(x)=2x, \quad H_2(x)=4x^2-2,\\ H_3(x) =& 8x^3-12x, \quad H_4(x)=16x^4-48x^2+12,\\ H_{n+1}(x) =& 2xH_n(x)-2nH_{n-1}(x), \end{aligned}$$

且 $\|H_n\|^2=2^n n!\pi^{1/2}$。

修正 Hermite 多项式与 Hermite 多项式的区别仅仅是权重函数由 e^{-x^2} 改成了 $w(x)=e^{-x^2/2}$，内积为

$$< f,g >= \int_{-\infty}^{\infty} f(x)g(x)e^{-x^2/2}\,dx,$$

整个序列可以递推计算如下：

$$\begin{aligned} He_0(x) =& 1, \quad He_1(x)=x, \quad He_2(x)=x^2-1,\\ He_3(x) =& x^3-3x, \quad He_4(x)=x^4-6x^2+3,\\ He_{n+1}(x) =& xHe_n(x)-nHe_{n-1}(x), \end{aligned}$$

且 $\|He_n\|^2=n!\sqrt{2\pi}$。

统计计算中常常会遇到没有解析表达式的函数的计算问题，比如，标准正态分布函数 $\Phi(x)=\int_{-\infty}^{x}\phi(u)du=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2}du$，我们可以用数值积分方法把 $\Phi(x)$ 计算到很高精度，但是这样计算会极为耗时。为此，我们可以用 Hermite 多项式或修正 Hermite 多项式逼近 $\Phi(x)$ 到一定的精度，计算多项式系数时可以用数值积分计算，这些系数计算可能耗时很多但是只要找到了逼近多项式就可以一劳永逸了。当然，经过多年研究，常见分布函数都已经有了很好的近似公式，比如

$$\Phi(x)=\begin{cases}\frac{1}{2}+\phi(x)\sum_{n=0}^{\infty}\frac{1}{(2n+1)!!}x^{2n+1}, & 0\le x\le 3,\\ 1-\phi(x)\sum_{n=0}^{\infty}(-1)^n\frac{(2n-1)!!}{x^{2n+1}}, & x>3\end{cases}\tag{4.7}$$

但是，在各种各样的统计模型中我们还是会遇到很多需要快速计算的函数，这些函数很可能没有高阶导数，每计算一次花费很长时间，我们可以用多项式逼近或插值方法给出近似公式，然后每次用近似公式计算函数值。

有些统计方法需要估计一个函数，这时可以用正交多项式级数表示这个函数，只要估计级数展开所需的系数。

4.1.2 连分式逼近

用正交多项式近似函数的好处是公式和计算简单，容易计算展开系数，容易微分、积分。但是多项式仅能逼近有限区间上的函数，容易出现很强的波动，尤其是在区间边界处, 而且多项式没有渐近线，没有无穷函数值的点，很多函数，尤其是取值区间包含正负无穷的函数，用正交多项式很难得到高精度。

有理函数能克服多项式的这些缺点。分子和分母都是多项式的分式叫做有理函数。逼近函数时，在相同参数个数条件下有理函数经常比多项式逼近好。有理函数可以从幂级数展开表示中解出。有理函数在计算时可以化为连分式来计算，连分式可以高效地计算。

连分式是像

$$1+\cfrac{1}{2+\cfrac{1}{3+\cfrac{1}{4}}}=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}$$

这样的分式，等号右侧是对左侧的分式的简写。一般地有

$$R_n=b_0+\cfrac{a_1}{b_1+\cfrac{a_2}{b_2+\ddots+\cfrac{a_n}{b_n}}}=b_0+\frac{a_1}{b_1}+\frac{a_2}{b_2}+\cdots+\frac{a_n}{b_n}.\tag{4.8}$$

计算连分式(4.8), 可以使用如下反向算法或正向算法，正向算法适用于无穷连分式的计算。

反向算法：

$z \leftarrow b_n$
for$(k = n, n-1, \ldots, 2, 1)$ {
  $z \leftarrow b_{k-1} + \frac{a_k}{z}$
}
输出 $R_n \leftarrow z_0$

正向算法：

$A_0 \leftarrow b_0$, $B_0 \leftarrow 1$
$A_1 \leftarrow b_1 b_0 + a_1$, $B_1 \leftarrow b_1$ # 效果是 $R_1 = A_1/B_1 = b_0 + \frac{a_1}{b_1}$
for$(k = 2, \ldots, n)$ {
  $A_k = b_k A_{k-1} + a_k A_{k-2}$
  $B_k = b_k B_{k-1} + a_k B_{k-2}$
}
输出 $R_n = A_n/B_n$，同时 $R_k = A_k/B_k, k = 1, 2, \ldots, n$

计算幂级数时，如果预先知道需要计算的项数，可以把幂级数化为无穷连分数，用连分数的正向算法计算:

$$\sum_{i=0}^{\infty} a_i x^i = \frac{b_0}{1} - \frac{b_1 x}{1 + b_1 x} - \frac{b_2 x}{1 + b_2 x} - \cdots \tag{4.9}$$

其中 $b_0 = a_0$, $b_i = a_i/a_{i-1} (i > 0)$.

有理分式可以化为连分式计算。转化的一般规则是:

- 如果分子分母都有常数项，如 $\frac{1 + 2x + x^3}{1 + 2x}$ 从分式中减去此常数项，使得分子以 x 为因子 (如 $1 + \frac{x^3}{1 + 2x}$)。
- 如果分子中有 x 因子而分母中有常数项，则把分式变成分子只有 x。(如 $\frac{x}{\frac{1 + 2x}{x^2}}$)
- 如果分母中有 x 因子而分子中有常数项 (如 $\frac{1 + 2x}{x^2}$)，写成倒数形式 (如 $\frac{1}{\frac{x^2}{1 + 2x}}$) 再处理。

- 转换的想法是分子上只留常数项或一次项。

例如

$$\begin{aligned}\frac{1+2x+x^3}{1+2x} =&\frac{1}{1}+\left(\frac{1+2x+x^3}{1+2x}-\frac{1}{1}\right) \text{ (减去都有的常数项)}\\ =&1+\frac{x^3}{1+2x}=1+\frac{x}{\frac{1+2x}{x^2}} \text{ (分子仅留 } x\text{)}\\ =&1+\frac{x}{\frac{1}{\frac{x^2}{1+2x}}} \text{ (分母有 } x \text{ 但分子有常数项时化为倒数)}\\ =&1+\frac{x}{\frac{1}{\frac{x}{2+\frac{1}{x}}}} \text{ (分子仅留 } x\text{)}\\ =&1+\frac{x}{0}+\frac{1}{0}+\frac{x}{2}+\frac{1}{x}.\end{aligned}$$

对于一般有理分式

$$g(x)=\frac{\sum_{i=0}^{\infty}a_{1i}x^i}{\sum_{i=0}^{\infty}a_{0i}x^i} \tag{4.10}$$

可化为

$$g(x)=\frac{a_{10}}{a_{00}}+\frac{a_{20}x}{a_{10}}+\frac{a_{30}x}{a_{20}}+\cdots \tag{4.11}$$

其中

$$a_{mi}=a_{m-1,0}a_{m-2,i+1}-a_{m-2,0}a_{m-1,i+1},\ i=0,1,2,\ldots;\ m=2,3,4,\ldots \tag{4.12}$$

有理分式化为连分式还可以用如下公式：

$$g(x)=\frac{\sum_{i=0}^{\infty}b_{1i}x^i}{\sum_{i=0}^{\infty}b_{0i}x^i}=\frac{1}{d_0}+\frac{x}{d_1}+\frac{x}{d_2}+\cdots \tag{4.13}$$

其中 $d_m=\frac{b_{m,0}}{b_{m+1,0}}$,

$$b_{m+2,i}=b_{m,i+1}-d_m b_{m+1,i+1},\ i=0,1,2,\ldots;m=0,1,2,\ldots \tag{4.14}$$

例如，计算标准正态分布函数的(4.7)可以用 (4.13)化为如下连分式：

$$\Phi(x)=\begin{cases}\frac{1}{2}+\frac{\phi(x)x}{1-\frac{x^2}{3+\frac{2x^2}{5-\cdots+\frac{(-1)^k kx^2}{2k+1+\cdots}}}}, & 0\le x\le 3,\\ 1-\frac{\phi(x)}{x+\frac{1}{x+\frac{2}{x+\cdots+\frac{k}{x+\cdots}}}}, & x>3\end{cases}. \tag{4.15}$$

4.1.3 逼近技巧

大多数计算机语言都提供了 $\sqrt{x}$、$\ln x$、e^x、$\sin x$ 等函数，R、SAS 等系统更提供了大多数概率密度、分布函数、分位数函数和随机数函数。我们应该尽量利用这些经过很多人验证的函数实现, 但是，也不能盲目相信已有的计算代码。

函数的数学定义、计算公式和计算机算法这三者是有很大差别的。计算机只能计算加减乘除，即使有些计算公式已经是泰勒展开式这样的形式，在实际编程计算时也要考虑浮点表示误差、误差积累等问题。另外，一个常用函数的算法需要同时满足精确性和高效率，算法的一点小改进就可能在反复调用中放大为很显著的效率改进。

以 $\sqrt{x}$ 计算为例，如果 y^2 的值和 x 差距在机器单位 U 以下，则可以用 y 作为 $\sqrt{x}$ 的值，尽管这样的 y 可能有多个。构造一个达到如此精度要求的高效率算法可能需要很多努力。我们知道，泰勒展开、多项式逼近等一般只适用于小范围而不适用于无穷区间，这时，“范围缩减”技术就可以帮我们把能够精确计算的范围扩大到全定义域。对 $\sqrt{x}$，我们可以用 $x_0 = 1$ 处的泰勒展开精确计算 $x \in [\frac{1}{2}, 1]$ 区间的 $\sqrt{x}$:

$$\sqrt{x} = 1 + \frac{1}{2}(x-1) - \frac{1}{8}(x-1)^2 + \frac{1}{16}(x-1)^3 - \frac{5}{128}(x-1)^4 + \cdots$$

实际设计算法时要考虑这个公式中正负项互相抵消和误差积累问题。用 $\sqrt{2} = \sqrt{4 \times \frac{1}{2}} = 2\sqrt{\frac{1}{2}}$ 计算出 $\sqrt{2}$，对于一般的 x，都可以写成

$$x = a2^k,\ a \in [\frac{1}{2}, 1]$$

从而

$$\sqrt{x} = \sqrt{a} \cdot 2^{k/2}$$

如果 k 是奇数，设 $k = 2m + 1$，则 $2^{k/2} = 2^m\sqrt{2}$。

4.2 插值

4.2.1 多项式插值

例 4.2.1. 下表是 $F(1, n)$ 分布的 0.95 分位数的部分值：

n	$\cdots$	20	$\cdots$	29	30	40	60	120	∞
$h(n)$	$\cdots$	4.35	$\cdots$	4.18	4.17	4.08	4.00	3.92	3.84

设关于函数 $h(n)$ 我们只有上述知识，但是还知道 $h(n)$ 具有一定光滑性。这里 $h(30)=4.17$, $h(40)=4.08$, 为了计算 $h(32)$，只要把 $(30,h(30))$ 和 $(40,h(40))$ 这两个点用线段连接，在 $(30,40)$ 之间用连接的线段作为 $h(x)$ 的近似值，就可以计算

$$h(32)=h(30)+\frac{h(40)-h(30)}{40-30}(32-30)\approx 4.15$$

称这样的近似计算为**线性插值**。

还可以进一步地近似。比如我们还知道 $h(20)=4.35$，则存在唯一的二次多项式 $f(x)$ 使得 $f(x)$ 穿过 $(20,h(20))$, $(30,h(30))$, $(40,h(40))$ 这三个点，用 $f(32)$ 来近似 $h(32)$，这叫做**抛物线插值**。

一般地，设已知函数 $h(x)$ 在点 $x_1,x_2,\ldots,x_n$ 上的值 $z_1,z_2,\ldots,z_n$ $(z_i=h(x_i))$，在假定 $h(x)$ 具有一定光滑度的条件下可以用函数 $f(x)$ 近似计算 $h(x)$ 在自变量 x 处的值。称 $h(x)$ 为**被插值函数**，$x_1,x_2,\ldots,x_n$ 叫做节点, $f(x)$ 叫做**插值函数**，一般使用多项式或分段多项式作插值函数。如果我们只能获得 $h(x)$ 在有限个点上的值，可以用插值方法近似计算 $h(x)$ 在其它自变量处的值。如果函数 $h(x)$ 每计算一个函数值到所需精度都要花费很长时间，可以预先在一些自变量处计算函数值并保存好，然后需要用到 $h(x)$ 的值时用插值法近似计算整个定义域上的函数值。如果 $h(x)$ 的计算精度要求不高，也可以预先计算并保存部分自变量处的函数值，然后用插值来近似函数值。插值得到的表达式可以用于积分、微分计算。

例4.2.1中用连接两点的线段作为插值公式的方法叫做线性插值。一般地，设已知 $(x_i,z_i),i=1,\ldots,n,\ z_i=h(x_i)$。对 $x\in[x_{i-1},x_i]$, 令

$$h(x)\approx f(x)=z_{i-1}+\frac{z_i-z_{i-1}}{x_i-x_{i-1}}(x-x_{i-1}) \tag{4.16}$$

(这就是直线方程的斜截式)。如果记 $\alpha=\frac{x-x_{i-1}}{x_i-x_{i-1}}$，即点 x 在插值区间 $[x_{i-1},x_i]$ 中前后位置比例，则

$$f(x)=(1-\alpha)z_{i-1}+\alpha z_i. \tag{4.17}$$

当 $h(x)$ 本身在 $[x_{i-1},x_i]$ 就是一次多项式时插值是精确的。

R 中用函数 `approx` 对给定的 n 个点在相邻点之间做线性插值。

当已知三个点的函数值时可以找到穿过这三个点的抛物线，用对应的二次多项式作为 $h(x)$ 的近似值。设 x_{-1},x_0,x_1 是三个点且距离为 $x_0-x_{-1}=x_1-x_0=d$，函数值分别为 z_{-1},z_0,z_1。对 $x\in[x_{-1},x_1]$，令 $\alpha=\frac{x-x_0}{d}$，则 $\alpha\in[-1,1]$, 令

$$f(\alpha)=\frac{z_{-1}-2z_0+z_1}{2}\alpha^2+\frac{z_1-z_{-1}}{2}\alpha+z_0, \tag{4.18}$$

则 $f(\alpha)$ 是 x 的二次函数，且经过 (x_{-1},z_{-1}), (x_0,z_0), (x_1,z_1) 这三个点, 可以用 $f(\alpha)$ 近似 $h(x)$ 的值。

如果有 n 个等距格点及对应函数值 $(x_i,z_i), i=1,2,\ldots,n$，$z_i=h(x_i)$, 为了用抛物线插值方法近似计算 $h(x)$, $x\in[x_1,x_n]$, 取 m_i 为区间 $[x_i,x_{i+1}]$ 的中点，对 $x\in[m_i,m_{i+1}]$ 用 (x_i,x_{i+1},x_{i+2}) 三点插值 $(m=2,3,\ldots,m_{n-2})$, 在 $[x_1,m_2]$ 内用 (x_1,x_2,x_3) 三个点插值，在 $[m_{n-2},x_n]$ 内用 x_{n-2},x_{n-1},x_n 三个点插值。

下面的定理保证了插值多项式的存在唯一性。

定理 4.2.1. 设有 $\{(x_i,z_i),i=1,2,\ldots,n\}$, $\{x_i,i=1,2,\ldots,n\}$ 互不相同，则存在唯一的不超过 $n-1$ 阶的多项式 $P_{n-1}(x)$ 使得

$$P_{n-1}(x_i)=z_i,\ i=1,2,\ldots,n.$$

证明. 设 $P_{n-1}(x)=a_0+a_1x+\ldots+a_{n-1}x^{n-1}$ 为多项式, 其中 $a_0,a_1,\ldots,a_{n-1}$ 为待定常数。为使 $P_{n-1}(x_i)=z_i$, $i=1,2,\ldots,n$，应有

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{n-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \tag{4.19}$$

注意到此方程组的系数矩阵行列式为 Vandermonde 行列式，行列式值为

$$\prod_{1\le i<j\le n}(x_j-x_i)\neq 0$$

所以方程组存在唯一解, 即存在唯一的次数不超过 $n-1$ 的多项式 $P_{n-1}(x)$ 使 $P_{n-1}(x)$ 通过 $(x_i,z_i),i=1,2,\ldots,n$ 这 n 个点。 □

由定理4.2.1的证明可见求插值多项式的方法之一是解线性方程组(4.19)，但是这样难以得到用初始值 $(x_i,z_i),i=1,2,\ldots,n$ 表示的函数表达式。

Lagrange 插值法

获得插值多项式表达式的一种方法是 Lagrange 插值法。令

$$d_i(x)=\frac{\prod\limits_{j\neq i}(x-x_j)}{\prod\limits_{j\neq i}(x_i-x_j)}$$

则 $d_i(x)$ 是 $n-1$ 阶多项式, 在各 $\{x_i\}$ 上满足:

$$\begin{cases} d_i(x_i) = 1 \\ d_i(x_k) = 0, k \neq i \end{cases}$$

令

$$f(x) = \sum_{i=1}^{n} z_i d_i(x) \tag{4.20}$$

则 $f(x)$ 是通过指定的 n 个点的不超过 $n-1$ 阶的多项式，根据定理4.2.1可知(4.20)给出的多项式是唯一的插值多项式。

为了用 Lagrange 方法插值，首先需要确定 n 个多项式 $d_i(x), i = 1, 2, \ldots, n$ 的系数。为了便于计算，记

$$g(x) = (x - x_1)(x - x_2)\cdots(x - x_n) = \prod_{i=1}^{n}(x - x_i), \tag{4.21}$$

则有

$$g'(x) = \sum_{k=1}^{n}\prod_{j\neq k}(x - x_j),$$
$$g'(x_i) = \sum_{k=1}^{n}\prod_{j\neq k}(x_i - x_j) = \prod_{j\neq i}(x_i - x_j)$$

于是可知

$$d_i(x) = \frac{g(x)/(x - x_i)}{g'(x_i)}. \tag{4.22}$$

这样，只要算出 $g(x)$ 的系数和 $g'(x)$ 的系数, 然后用多项式除法写出 $g(x)/(x - x_i)$ 的系数，就可以用(4.22)得到 $d_i(x)$ 的系数。计算插值多项式 $f(x)$ 的系数的算法如下：

写出 $g(x)$ 的展开式 $g(x) = \sum_{i=0}^{n} a_i x^i$.
写出 $g'(x)$ 的展开式 $g'(x) = \sum_{i=0}^{n-1}(i+1)a_{i+1}x^i$.
计算各 $g'(x_i)$.
写出各 $g(x)/(x - x_i)$ 的展开式.
用下式给出插值多项式表达式：
$f(x) = \sum_{i=1}^{n} \frac{z_i}{g'(x_i)}[g(x)/(x - x_i)]$

上述算法中用到了多项式加减法、乘法和除法。

为计算多项式乘法 $\sum_{i=0}^{n} a_i x^i \sum_{j=0}^{m} b_j x^j$, 记 $a_i = 0 (i > n)$, $b_j = 0 (j > m)$, 有

$$\begin{aligned}
&\sum_{i=0}^{n} a_i x^i \sum_{j=0}^{m} b_j x^j = \sum_{i=0}^{n} \sum_{j=0}^{m} a_i b_j x^{i+j} \ (\text{令 } k = i + j) \\
=&\sum_{i=0}^{n} \sum_{k=i}^{i+m} a_i b_{k-i} x^k = \sum_{k=0}^{m+n} \left(\sum_{i=0}^{k} a_i b_{k-i} \right) x^k \\
=&\sum_{k=0}^{m+n} \left(\sum_{i=\max(0,k-m)}^{\min(k,n)} a_i b_{k-i} \right) x^k \ \ (\text{注意 } 0 \le i \le n,\ 0 \le k - i \le m) \\
=&\sum_{k=0}^{m+n} c_k x^k,
\end{aligned}$$

乘积多项式的系数为

$$c_k = \sum_{i=\max(0,k-m)}^{\min(k,n)} a_i b_{k-i},\ k = 0, 1, \ldots, m + n. \tag{4.23}$$

$\{c_k\}$ 是 $\{a_k\}$ 和 $\{b_k\}$ 的卷积。如果 $\{a_k\}$, $\{b_k\}$ 是两个数列，称

$$c_k = \sum_{i=-\infty}^{\infty} a_i b_{k-i},\ k \in \mathbb{Z}$$

($\mathbb{Z}$ 为所有整数) 为 $\{a_k\}$, $\{b_k\}$ 的卷积。

注: R 函数 `filter` 和 `convolve` 可以计算两个有限序列的卷积。

关于多项式插值的误差有如下定理。

定理 4.2.2. 设 $h(x) \in C^{n-1}[a, b]$，$h^{(n)}(x)$ 在 (a, b) 内存在，设 $\{x_i, i = 1, 2, \ldots, n\}$ 互不相同，$g(x) = \prod_{i=1}^{n} (x - x_i)$, $\{x_i, i = 1, 2, \ldots, n\}$ 的最小值到最大值的区间为 I，则对 $\forall x \in I$, 存在 $\xi \in I$ 使

$$R(x) = h(x) - f(x) = \frac{h^{(n)}(\xi)}{n!} g(x). \tag{4.24}$$

证明. 不妨设 $x_1, x_2, \ldots, x_n$ 从小到大排列。对 $x \in \{x_1, x_2, \ldots, x_n\}$, 显然 $R(x) = 0$。对给定的 $x \in [a, b] \backslash \{x_1, x_2, \ldots, x_n\}$, 定义辅助函数

$$H(t) = R(t) - \frac{g(t)}{g(x)} R(x)$$

易见 $H(x_i)=0, i=1,2,\ldots,n$，且 $H(x)=0$，所以 $H(t)$ 有 $n+1$ 个相异零点。由 Rolle 定理，$H(t)$ 的每两个相邻零点中间必有一个点导数等于零，所以 $H^{(1)}(t)$ 至少有 n 个相异零点。类似讨论可知 $H^{(j)}(t)$ 至少有 $n+1-j$ 个相异零点，$j=1,2,\ldots,n$。因此 $H^{(n)}(t)$ 至少有一个零点。设 ξ 是 $H^{(n)}(t)$ 的一个零点，则

$$\begin{aligned}0=&H^{(n)}(\xi)=R^{(n)}(\xi)-\frac{g^{(n)}(\xi)}{g(x)}R(x)\\=&h^{(n)}(\xi)-f^{(n)}(\xi)-\frac{n!}{g(x)}R(x)\\=&h^{(n)}(\xi)-\frac{n!}{g(x)}R(x)\end{aligned}$$

于是

$$R(x)=\frac{h^{(n)}(\xi)}{n!}g(x)$$

定理得证。 □

$g(x)$ 在靠近 $\{x_i\}$ 最小值和最大值的位置绝对值较大，说明多项式插值在边界处误差较大, 见图4.1的左下图。所以，多项式插值并不是多项式的阶数越高越好，阶数太高的多项式的曲线形状会变得很复杂，插值误差反而会很大。

为了使(4.24)中的 $g(x)$ 不要太大，还可以考虑适当选取 x_i 的位置。对 $(-1,1)$ 区间，取 x_i 为 Chebyshev I 型多项式 $T_n(x)$ 的零点 $x_i=\cos\left(\frac{2i-1}{2n}\pi\right)$ $(i=1,2,\ldots,n)$ 可以使 $g(x)$ 最小。

例 4.2.2 (用 Lagrange 法进行二次插值). 设已知 $(x_i,z_i), i=1,2,3$ 三个点，求通过这三个点的抛物线方程。按照 Lagrange 方法，令

$$\begin{aligned}u_1(x)=&(x-x_2)(x-x_3)=x^2-(x_2+x_3)x+x_2x_3\\u_2(x)=&(x-x_1)(x-x_3)=x^2-(x_1+x_3)x+x_1x_3\\u_3(x)=&(x-x_1)(x-x_2)=x^2-(x_1+x_2)x+x_1x_2\end{aligned}$$

则 $d_i(x)=u_i(x)/u_i(x_i)$。记 $b_i=z_i/u_i(x_i)$ 则 $f(x)=\sum_{i=1}^3 b_iu_i(x)$, 于是抛物线插值公式为

$$\begin{aligned}f(x)=&(b_1+b_2+b_3)x^2\\&-\left(b_1(x_2+x_3)+b_2(x_1+x_3)+b_3(x_1+x_2)\right)x\\&+\left(b_1x_2x_3+b_2x_1x_3+b_3x_1x_2\right)\end{aligned}\tag{4.25}$$

例如，对 $h(x)=\sqrt{x}$ 在 $x=1/16,1/4,1$ 用抛物线插值，有

$$b_1=\frac{1/4}{u_1(\frac{1}{16})}=\frac{64}{45},\ b_2=\frac{1/2}{u_2(\frac{1}{4})}=-\frac{32}{9}=-\frac{160}{45},\ b_3=\frac{1}{u_3(1)}=\frac{64}{45}$$

于是插值函数

$$\begin{aligned}h(x)=&\left(\frac{64}{45}-\frac{160}{45}+\frac{64}{45}\right)x^2\\&-\left(\frac{64}{45}(\frac{1}{4}+1)-\frac{160}{45}(\frac{1}{16}+1)+\frac{64}{45}(\frac{1}{16}+\frac{1}{4})\right)x\\&+\left(\frac{64}{45}\frac{1}{16}\frac{1}{4}-\frac{160}{45}\frac{1}{16}\cdot 1+\frac{64}{45}\frac{1}{4}\cdot 1\right)\\=&\frac{1}{45}(7+70x-32x^2)\end{aligned}$$

在 $[0,1]$ 插值的平均绝对误差约为 0.033。

在 $(-1,1)$ 区间按照 Chebyshev I 型零点取 x_i 为 $-\sqrt{3}/2,0,\sqrt{3}/2$，于是在 $[0,1]$ 区间取 x_i 分别为 $\frac{2-\sqrt{3}}{2},\frac{1}{2},\frac{2+\sqrt{3}}{2}$, 这时在 $[0,1]$ 区间插值的平均绝对误差约为 0.015。

也可以考虑用有理函数插值。设

$$f(x)=\frac{a_0+a_1x}{1+b_1x},$$

令 $f(x_i)=z_i,\ i=1,2,3$，仍取 x_i 为 $1/16,1/4,1$, 可得方程

$$\begin{aligned}a_0+\frac{1}{16}a_1=&\frac{1}{4}\left(1+\frac{1}{16}b_1\right)\\a_0+\frac{1}{4}a_1=&\frac{1}{2}\left(1+\frac{1}{4}b_1\right)\\a_0+a_1=&1\cdot(1+b_1)\end{aligned}$$

解得

$$f(x)=\frac{\frac{1}{7}+2x}{1+\frac{8}{7}x},$$

这时在 $[0,1]$ 区间插值的平均绝对误差约为 0.014。

牛顿差商公式 *

牛顿差商公式是另外一种给出插值多项式表达式的方法。对函数 $h(x)$，分点 $x_i, i = 1, \ldots, n$，定义各阶差商为：

$$\begin{aligned}
h[x_i] =& h(x_i) \\
h[x_i, x_j] =& \frac{h(x_j) - h(x_i)}{x_j - x_i} \\
h[x_i, x_j, x_k] =& \frac{h[x_j, x_k] - h[x_i, x_j]}{x_k - x_i} \\
& \cdots\cdots\cdots \\
h[x_1, x_2, \ldots, x_n] =& \frac{h[x_2, x_3, \ldots, x_n] - h[x_1, x_2, \ldots, x_{n-1}]}{x_n - x_1}
\end{aligned}$$

则 $h[x_1, x_2, \ldots, x_n]$ 关于自变量对称，即以任意自变量次序计算 $h[x_1, x_2, \ldots, x_n]$ 结果都相同。另外，在定理4.2.2条件下存在 $\xi \in [\min\{x_i\}, \max\{x_i\}]$ 使得

$$h[x_1, x_2, \ldots, x_n] = \frac{h^{(n-1)}(\xi)}{(n-1)!}.$$

利用牛顿差商公式也可以得到多项式插值公式

$$\begin{aligned}
h(x) =& P_{n-1}(x) + R_n(x) \\
P_{n-1}(x) =& h(x_1) + (x - x_1)h[x_1, x_2] \\
& + (x - x_1)(x - x_2)h[x_1, x_2, x_3] \\
& + \cdots \\
& + (x - x_1)(x - x_2)\cdots(x - x_{n-1})h[x_1, x_2, \ldots, x_n] \qquad (4.26) \\
R_n(x) =& (x - x_1)(x - x_2)\cdots(x - x_{n-1})(x - x_n) \\
& \cdot h[x_1, x_2, \ldots, x_n, x] \qquad (4.27)
\end{aligned}$$

其中 $P_{n-1}(x)$ 是 $n-1$ 次插值多项式，根据定理4.2.1可知 $P_{n-1}(x)$ 与 Lagrange 插值多项式相同。$R_n(x)$ 为余项，关于余项估计有:

$$h[x_1, x_2, \ldots, x_n, x] = \frac{h^{(n)}(\eta)}{n!},$$

其中 $\eta \in [\min(x_1, x_2, \ldots, x_n, x), \max(x_1, x_2, \ldots, x_n, x)]$。

4.2.2 样条插值介绍

如果有 n 个点 $\{(x_i, z_i), i = 1, 2, \ldots, n\}$ 要插值，当 n 较大时，用 $n-1$ 阶插值多项式可以通过这些点，但是得到的曲线会过于复杂, 与真实值差距很大，如果用插值函数计算 $\{(x_i, z_i), i = 1, 2, \ldots, n\}$ 所在区间外的点则误差更大。如果仅仅用线段连接或者用抛物线连接，那么在两段交界的地方又不够光滑。图4.1用几种方法对函数 $h(x) = 1/(1 + x^2)$ 在 $x \in [-4, 4]$ 进行了插值，只用到了 7 个已知点。从例图可以看出，线性插值可能在连接两段的地方形成尖角转弯，抛物线插值在两段连接的地方也有明显转折，用通过给定的 7 个点的 6 阶多项式插值，出现了很大的波折，与真实函数差距较大。**样条插值**也是用分段的多项式进行插值，但样条插值保证连接处是光滑的，同时构成的曲线又不会太复杂。

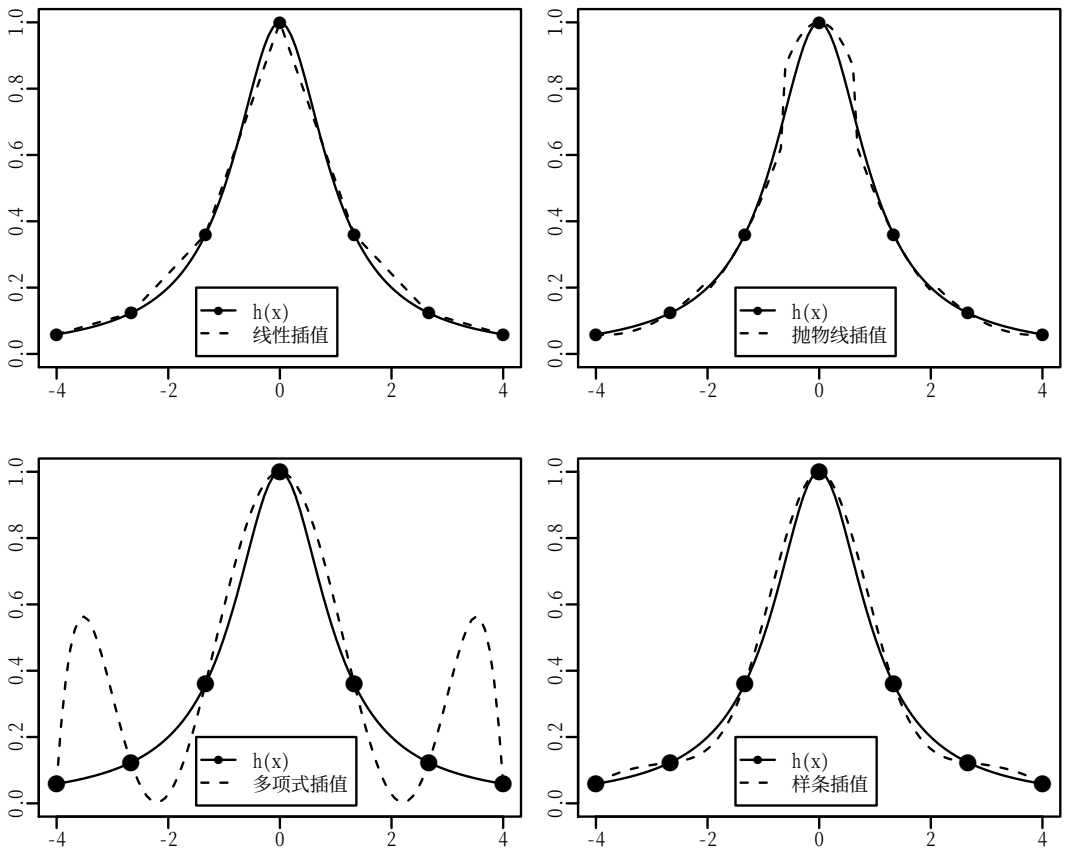

图 4.1: 函数 $1/(1 + x^2)$ 的线性插值、抛物线插值、$n-1$ 阶多项式插值和样条插值

设已知 n 个点 $\{(x_i,z_i),i=1,\dots,n\}$，$x_1<x_2<\cdots<x_n$, 求有二阶连续导数的 $S(x)$ 使得 $S(x)$ 通过这 n 个点，即求 $S(x)$ 使

$$\begin{cases} S(x_i)=z_i, i=1,\dots,n \\ \min\limits_{S}\int |S''(x)|^2dx \end{cases}$$

满足条件的函数就是三次样条函数。三次样条函数是分段函数，以各 x_i 为分界点，称 $\{x_1,x_2,\dots,x_n\}$ 为结点 (knots)，三次样条函数 $S(x)$ 在每一段内为三次多项式，其一阶导数 $S'(x)$ 连续，为分段二次多项式，其二阶导数 $S''(x)$ 连续，为分段线性函数。

样条插值优点是保证了曲线光滑，可以对比较光滑的各种形状的函数进行插值同时插值曲线不会太复杂，得到的函数表达式可以用于数值积分和数值微分。

为了求样条函数 $S(x)$ 的每一段的表达式，记 $d_j=x_j-x_{j-1}, j=2,\dots,n$。因为 $S''(x)$ 分段线性，可记 $S''(x_j)=M_j$，用线性插值公式，对 $x\in[x_{j-1},x_j]$ 有

$$S''(x)=\frac{M_{j-1}(x_j-x)}{d_j}+\frac{M_j(x-x_{j-1})}{d_j},$$

积分得

$$\begin{aligned} S'(x)=&c_1+(-1)\frac{M_{j-1}(x_j-x)^2}{2d_j}+\frac{M_j(x-x_{j-1})^2}{2d_j}, \\ S(x)=&c_1x+c_2+\frac{M_{j-1}(x_j-x)^3}{6d_j}+\frac{M_j(x-x_{j-1})^3}{6d_j}, \end{aligned}$$

改写为

$$S(x)=\frac{M_{j-1}(x_j-x)^3+M_j(x-x_{j-1})^3}{6d_j}+c_1'\frac{x_j-x}{d_j}+c_2'\frac{x-x_{j-1}}{d_j}.$$

根据 $S(x_{j-1})=z_{j-1}$ 和 $S(x_j)=z_j$ 解出 c_1',c_2'，有

$$\begin{aligned} S(x)=&\frac{M_{j-1}(x_j-x)^3+M_j(x-x_{j-1})^3}{6d_j} \\ &+\left(z_{j-1}-\frac{M_{j-1}d_j^2}{6}\right)\frac{x_j-x}{d_j}+\left(z_j-\frac{M_jd_j^2}{6}\right)\frac{x-x_{j-1}}{d_j}, \\ &x\in[x_{j-1},x_j]. \end{aligned}$$

只要求出 $M_j,j=1,\dots,n$，则三次样条插值函数 $S(x)$ 在每一段的表达式就完全确定。

用 $S'(x_j-0)=S'(x_j+0), j=2,\dots,n-1$ 可以得到 $\{M_j\}$ 的 $n-2$ 个线性方程，每一方程只涉及 M_{j-1},M_j,M_{j+1}. 再增加两个线性限制条件可以解出 $\{M_j,j=1,\dots,n\}$。比如，加

限制条件 $M_1 = 0, M_n = 0$，即在边界处函数为线性函数，这样解出的样条插值函数称为**自然样条**。

R 中样条插值函数为 `spline(x, y, n, method="natural")` , 结果为包含元素 `x` 和 `y` 的列表。

样条函数除了可以用于插值，还可以用于平滑点 $(x_i, z_i), i = 1, 2, \ldots, n$。插值函数需要穿过所有已知点,而平滑则只要与已知点很接近就可以了。用样条函数在点 $(x_i, z_i), i = 1, 2, \ldots, n$ 处进行平滑叫做样条回归，结果是以 $x_1, x_2, \ldots, x_n$ 为结点的三次样条函数，但在 x_i 处的值不需要等于 z_i。样条回归是使得

$$Q(S) = \sum_{i=1}^{n}(z_i - S(x_i))^2 + \lambda n \int |S''(x)|^2 dx$$

最小的函数，其中 $\lambda > 0$ 是代表光滑程度的参数，λ 越大得到的样条函数越光滑。

样条函数用作回归函数时也可以不取已知点 $\{(x_i, z_i), i = 1, 2, \ldots, n\}$ 的横坐标为结点，而是单独取结点 $k_1, k_2, \ldots, k_m$，一般 m 比 n 小得多，结点 k_j 作为回归关系改变的分界点。

4.3 数值积分和数值微分

4.3.1 数值积分的用途

在 §3.2给出了用随机模拟积分的方法。随机模拟积分在高维或者积分区域不规则时有优势，在低维、积分区域规则且被积函数比较光滑时用数值积分方法可以得到更高精度，也不会引入随机误差。

积分、最优化、矩阵计算都是在统计问题中最常见的计算问题，在统计计算中经常需要计算积分。比如，从密度 $p(x)$ 计算分布函数 $F(x)$，如果没有解析表达式和精确的计算公式，需要用积分来计算：

$$F(x) = \int_{-\infty}^{x} p(u)du$$

用积分给出部分函数值后可以用插值和函数逼近得到 $F(x)$ 的近似公式。

已知联合密度 $p(\boldsymbol{x}_1, \boldsymbol{x}_2)$ 要求边缘密度 $p(\boldsymbol{x}_1)$，要用积分计算

$$p(\boldsymbol{x}_1) = \int p(\boldsymbol{x}_1, \boldsymbol{x}_2)d\boldsymbol{x}_2.$$

贝叶斯分析的主要问题是已知先验密度 $\pi(\theta)$ 和似然函数 $p(\boldsymbol{x}|\theta)$ 后求后验密度 $p(\theta|\boldsymbol{x})$:

$$p(\theta|\boldsymbol{x}) = \frac{p(\theta, \boldsymbol{x})}{p(\boldsymbol{x})} = \frac{\pi(\theta)p(\boldsymbol{x}|\theta)}{\int \pi(u)p(\boldsymbol{x}|u)du}$$

大多数情况下不能得到后验密度 $p(\theta|\boldsymbol{x})$ 的解析表达式，也可能需要计算积分, 用后验密度求期望、平均损失函数也需要计算积分。

4.3.2 一维数值积分

数值积分的最简单方法是直接用达布和计算。更精确的积分方法是对被积函数进行多项式逼近然后对近似多项式用代数方法求积分。这些近似多项式可以不依赖于被积函数，只需要用被积函数的若干值。多项式逼近可以是在全积分区间上进行，也可以把积分区间分为很多小区间在小区间上逼近。分为小区间的方法适用性更好。

计算达布和的积分方法

为求定积分

$$I = \int_a^b f(x)dx, \tag{4.28}$$

把区间 $[a,b]$ 均匀分为 n 段，分点为 $x_0 = a, x_1, \ldots, x_{n-1}, x_n = b$，间隔为 $h = (b-a)/n$, 可以用

$$D_n = \sum_{i=1}^{n} f(x_i)h \tag{4.29}$$

近似计算 I。把 D_n 改写成

$$D_n = \sum_{i=0}^{n-1} f(x_{i+1})h$$

可以看出相当于把曲边梯形面积

$$\int_{x_i}^{x_{i+1}} f(x)dx \tag{4.30}$$

用以 $[x_i, x_{i+1}]$ 为底、右侧高度 $f(x_{i+1})$ 为高的矩形面积近似 (见图4.2), (4.29)中的 $f(x_i)$ 是小区间右端点的函数值。此方法容易理解，但用区间端点近似整个区间的函数值误差较大，精度不好，基本不使用公式(4.29)计算一元函数积分。如果使用区间中点作为代表，公式变成

$$M_n = h\sum_{i=0}^{n-1} f\left(a + \left(i + \frac{1}{2}\right)h\right) \quad (h = \frac{b-a}{n}) \tag{4.31}$$

这个公式称为中点法则，余项为

$$R_n = I - M_n = \frac{(b-a)^3}{24n^2} f''(\xi),\ \xi \in [a,b], \tag{4.32}$$

当 $f''(x)$ 有界时中点法则的精度为 $O(n^{-2})$，精度比(4.29)高得多，与下面的梯形法则精度相近。

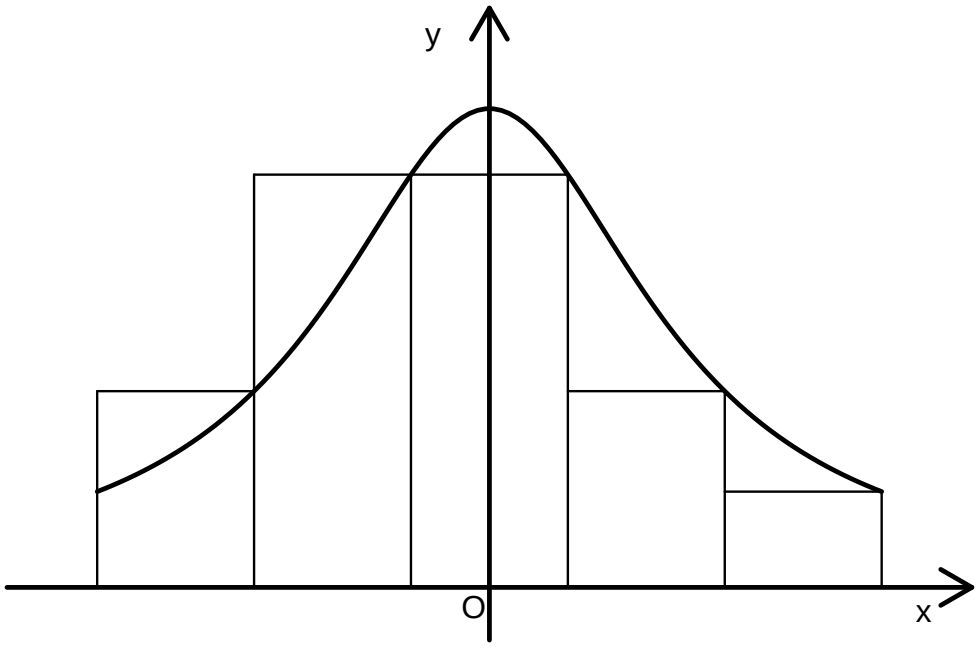

图 4.2: 达布和数值积分图示

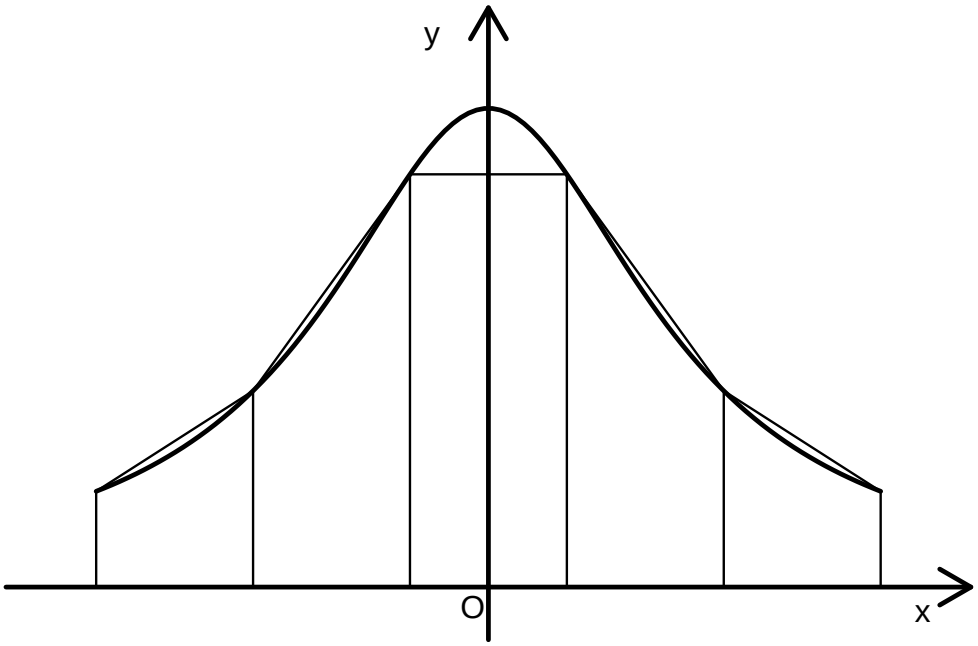

图 4.3: 梯形法则数值积分图示

梯形法则

在小区间 $[x_i, x_{i+1}]$ 内不是用常数而是用线性插值代替 $f(x)$，即用梯形代替曲边梯形 (参见图4.3)，可以得到如下的积分公式

$$\begin{aligned} f(x) \approx & f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{h}(x - x_i), \ x \in [x_i, x_{i+1}] \\ \int_{x_i}^{x_{i+1}} f(x)dx \approx & \frac{f(x_i) + f(x_{i+1})}{2}h \end{aligned} \tag{4.33}$$

$$\begin{aligned} \int_a^b f(x)\,dx = & \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,dx \\ \approx & \frac{h}{2}\left\{ f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b) \right\} \triangleq T_n \end{aligned} \tag{4.34}$$

余项为

$$R_n = I - T_n = -\frac{(b-a)^3}{12n^2} f''(\xi), \xi \in (a,b). \tag{4.35}$$

(4.34)叫做复合梯形公式，在 $f''(x)$ 有界时算法精度为 $O(n^{-2})$。

辛普森 (Simpson) 法则

在等距的 (x_{-1}, x_0, x_1) 的区间中用抛物线插值公式近似 $f(x)$:

$$\begin{aligned} f(x) \approx & \frac{1}{2}\left(f(x_{-1}) - 2f(x_0) + f(x_1)\right)\left(\frac{x-x_0}{h/2}\right)^2 \\ & + \frac{1}{2}\left(f(x_1) - f(x_{-1})\right)\left(\frac{x-x_0}{h/2}\right) + f(x_0) \end{aligned} \tag{4.36}$$

其中 $h = x_1 - x_{-1} = 2(x_0 - x_{-1})$, 积分得

$$\int_{x_{-1}}^{x_1} f(x)\,dx = \frac{h}{6}\left\{f(x_{-1}) + 4f(x_0) + f(x_1)\right\}, h = t_1 - t_{-1}. \tag{4.37}$$

把区间 $[a,b]$ 等分为 n 份, 记 $h = (b-a)/n$, 记 $x_i = a + ih$, $i = 0, 1, 2, \ldots, n$, 则 $x_0, x_1, \ldots, x_n$ 把 $[a,b]$ 等分为 n 份，然后在每个小区间内取中点，记为 $x_{i+\frac{1}{2}} = a + (i+\frac{1}{2})h$, $i = 0, 1, \ldots, n-1$, 在 $[x_i, x_{i+1}]$ 内用公式(4.37)并把 n 个小区间的积分相加，得

$$\begin{aligned} I = & \int_a^b f(x)\,dx \\ \approx & \frac{h}{6}\left(f(a) + 4\sum_{i=0}^{n-1} f(x_{i+\frac{1}{2}}) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right) \triangleq S_n, \end{aligned} \tag{4.38}$$

余项为

$$R_n = I - S_n = -\frac{(b-a)^5}{2880n^4} f^{(4)}(\xi), \xi \in (a,b). \tag{4.39}$$

(4.38)叫做复合辛普森公式，在 $f^{(4)}(x)$ 有界时复合辛普森公式误差为 $O(n^{-4})$。

牛顿-柯蒂斯 (Newton-Cotes 公式)*

公式(4.33)和(4.37)分别是在小区间用线性插值和抛物线插值近似被积函数 $f(x)$ 得到的积分公式。一般地，对等距节点 $(x_0 = a, x_1, \ldots, x_n = b)$ 可以进行 Lagrange 插值，得到 n 阶插值多项式 $P_n(x)$:

$$P_n(x) = \sum_{j=0}^{n} f(x_j) \frac{\prod_{k\neq j}(x - x_k)}{\prod_{k\neq j}(x_j - x_k)}$$

用 $P_n(x)$ 在 $[a,b]$ 上的积分近似 $\int_a^b f(x)dx$，有

$$\int_a^b f(x)\,dx \approx \int_a^b P_n(x)\,dx \tag{4.40}$$

$$= \sum_{j=0}^{n} \left\{ \int_a^b \frac{\prod_{k\neq j}(x-x_k)}{\prod_{k\neq j}(x_j-x_k)} dx \right\} f(x_j) \tag{4.41}$$

注意到 $x_j = a+jh$, $h=(b-a)/n$, 做变量替换 $x=a+th$, 上式变成

$$\int_a^b f(x)\,dx \approx \frac{b-a}{n} \sum_{j=0}^{n} \left\{ \int_0^n \frac{\prod_{k\neq j}(t-k)}{\prod_{k\neq j}(j-k)} dt \right\} f(x_j)$$

$$= (b-a) \sum_{j=0}^{n} C_j^{(n)} f(x_j) \tag{4.42}$$

这是被积函数在节点上函数 $f(x_i)$ 值的线性组合，其中各线性组合系数 $C_j^{(n)}$ 不依赖于 f 和积分区间:

$$C_j^{(n)} = \frac{1}{n}\int_0^n \frac{\prod_{k\neq j}(t-k)}{\prod_{k\neq j}(j-k)} dt = \frac{1}{n}\frac{(-1)^{n-j}}{j!(n-j)!}\int_0^n \prod_{k\neq j}(t-k)dt. \tag{4.43}$$

公式(4.42)称为牛顿-柯蒂斯 (Newton-Cotes) 公式, 公式(4.33)和(4.37)分别是 $n=2$ 和 $n=3$ 的特例。一些 $C_j^{(n)}$ 的值见表4.2。$n=4$ 的牛顿-柯蒂斯公式叫做柯蒂斯法则，积分公式为

$$I = \int_a^b f(x)dx \approx \frac{b-a}{90}\left\{7f(x_0)+32f(x_1)+12f(x_2)+32f(x_3)+7f(x_4)\right\}, \tag{4.44}$$

$$x_i = a + i\cdot\frac{b-a}{4},\ i=0,1,2,3,4. \tag{4.45}$$

复合柯蒂斯公式把 $[a,b]$ 等分为 n 份，令 $h=(b-a)/n$, $x_i=a+ih, i=0,1,\ldots,n$, 然后把每个小区间 $[x_i,x_{i+1}]$ 等分为 4 份，内部分点为 $x_{i+\frac{1}{4}}=x_i+\frac{1}{4}h$, $x_{i+\frac{1}{2}}=x_i+\frac{1}{2}h$, $x_{i+\frac{3}{4}}=x_i+\frac{3}{4}h$, $i=0,1,\ldots,n-1$，在每个小区间 $[x_i,x_{i+1}]$ 内用(4.45)积分然后相加，得复合柯蒂斯公式

$$\int_a^b f(x)dx \approx \frac{h}{90}\left\{7f(a)+32\sum_{i=0}^{n-1} f(x_{i+\frac{1}{4}})+12\sum_{i=0}^{n-1} f(x_{i+\frac{1}{2}})\right.$$

$$\left.+32\sum_{i=0}^{n-1} f(x_{i+\frac{3}{4}})+14\sum_{i=1}^{n-1} f(x_i)+7f(b)\right\} \tag{4.46}$$

余项为

$$R = -\frac{(b-a)^7}{1013760n^6} f^{(6)}(\xi),\ \xi\in(a,b).$$

表 4.2: 牛顿-柯蒂斯积分系数

n	$C_j^{(n)}$
1	$\frac{1}{2}$, $\frac{1}{2}$
2	$\frac{1}{6}$, $\frac{2}{3}$, $\frac{1}{6}$
3	$\frac{1}{8}$, $\frac{3}{8}$, $\frac{3}{8}$, $\frac{1}{8}$
4	$\frac{7}{90}$, $\frac{32}{90}$, $\frac{12}{90}$, $\frac{32}{90}$, $\frac{7}{90}$
5	$\frac{19}{288}$, $\frac{75}{288}$, $\frac{50}{288}$, $\frac{50}{288}$, $\frac{75}{288}$, $\frac{19}{288}$
6	$\frac{41}{840}$, $\frac{216}{840}$, $\frac{27}{840}$, $\frac{272}{840}$, $\frac{27}{840}$, $\frac{216}{840}$, $\frac{41}{840}$

虽然我们可以得到高阶的牛顿-柯蒂斯公式，但是高阶的插值多项式一般在边界处会有很大误差，实际中我们一般最多用到 4 阶，然后用复合方法把 $[a,b]$ 划分为小区间，在小区间内用低阶的牛顿-柯蒂斯公式。

数值积分的代数精度 *

为计算积分

$$I=\int_a^b f(x)dx,$$

可以在 $[a,b]$ 的 n 个节点 $a\le x_1<x_2<\cdots<x_n\le b$ 上对 $f(x)$ 用 $n-1$ 阶多项式进行插值，用插值多项式的积分近似积分 I。由 Lagrange 插值公式(4.20)得

$$I\approx\sum_{j=1}^{n}\left\{\int_a^b\frac{\prod_{k\neq j}(x-x_k)}{\prod_{k\neq j}(x_j-x_k)}dx\right\}f(x_j) \tag{4.47}$$

$$=\sum_{j=1}^{n}A_jf(x_j), \tag{4.48}$$

其中 $\{A_j\}$ 不依赖于被积函数 $f(x)$。

在理想情况下(4.48)可以有较高精度。如果某插值方法对不超过 $n-1$ 次的多项式可精确表示而不能精确表示 n 次多项式，则称该插值方法有 $n-1$ 次**代数精度**。(4.48)若对不超过 $n-1$ 次的多项式积分可以得到准确值而 n 次则不行，则称此数值积分方法具有 $n-1$ 次代数精度。(4.48)至少具有 $n-1$ 次代数精度, 所以牛顿-柯蒂斯积分公式(4.42) 至少具有 n 次代数精度。

高斯-勒让德 (Gauss-Legendre) 积分公式 *

公式(4.48)至少有 $n-1$ 次代数精度。在(4.48)中适当选取节点 $\{x_j\}$ 的位置可以得到更高的代数精度, n 个节点可以达到 $2n-1$ 次代数精度, 还可以在无穷区间上积分。这样的数值积分方法叫做高斯型数值积分。

对至多 $2n-1$ 次多项式 $f(x)$，取 n 个节点 $\{x_k\}$ 为 n 次 Legendre 正交多项式 $P_n(x)$ 的 n 个零点，即

$$P_n(x) = c(x-x_1)\cdots(x-x_n),$$

则

$$f(x) = q(x)P_n(x) + r(x)$$

其中 $q(x)$ 和 $r(x)$ 都是至多 $n-1$ 次的多项式。由正交多项式性质, $P_n(x)$ 与任意至多 $n-1$ 次多项式都正交，于是

$$\int_{-1}^{1} q(x)P_n(x)\,dx = 0.$$

故

$$\int_{-1}^{1} f(x)\,dx = \int_{-1}^{1} r(x)\,dx,$$

而 $r(x)$ 为至多 $n-1$ 阶，其插值方法的积分是精确的，即

$$\int_{-1}^{1} f(x)\,dx = \int_{-1}^{1} r(x)\,dx = \sum_{k=1}^{n} A_k r(x_k)$$

等式成立，注意

$$P_n(x_k) = 0, k = 1, 2, \ldots, n,$$

所以 $r(x_k) = f(x_k)$, 于是

$$\int_{-1}^{1} f(x)\,dx = \sum_{k=1}^{n} A_k f(x_k)$$

即当 $f(x)$ 为至多 $2n-1$ 次多项式时这样选取节点的(4.48) 是精确的。可预先把 x_k 和 A_k 做表。

对于其他的积分区间及权函数 $w(x)$，要计算

$$I = \int_a^b f(x)w(x)\,dx,$$

也可以可以利用相应的正交多项式及零点解决。比如 Gauss-Laguerre 积分

$$\int_0^\infty f(x)e^{-x}\,dx$$

可以利用 Laguerre 多项式的零点作为节点进行插值，Gauss-Hermite 积分

$$\int_{-\infty}^\infty f(x)e^{-x^2}\,dx$$

可以利用 Hermite 多项式的零点作为节点进行插值。这样可以计算无穷区间上的积分。

高斯型积分适用于函数足够光滑，需要反复计算积分但每次计算被积函数值都耗时很多的情况，不适用于一般的被积函数。

变步长积分

对一般的函数，提高插值多项式阶数并不能改善积分精度，复合方法如 (4.34)和 (4.38)则取点越多精度越高。因为很难预估需要的点数，所以我们可以逐步增加取点个数直至达到需要的积分精度。一种节省取点个数的方法是每次取点仅增加原来小区间的中点，这样点数每次增加一倍但是仅需要再多计算一半点上的函数值。

比如，使用复合梯形公式(4.34):

$$\begin{aligned} T_n =& \frac{d}{2}\left\{f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right\}, \\ & x_i = a + id,\ i = 0, 1, \ldots, n,\ d = \frac{b-a}{n}, \end{aligned}$$

变步长为 $d/2$，节点增加到 $2n+1$ 个，原来的一个小区间 $[x_i, x_{i+1}]$ 中又增加了一个分点 $x_{i+\frac{1}{2}} = x_i + \frac{d}{2}$，原来在 $[x_i, x_{i+1}]$ 的梯形法则积分结果为

$$I_1 = \frac{d}{2}\left\{f(x_i) + f(x_{i+1})\right\},$$

增加 $x_{i+\frac{1}{2}}$ 节点后在 $[x_i, x_{i+1}]$ 的积分变成 $x_{i+\frac{1}{2}}$ 前后两半的积分之和，为

$$\begin{aligned} I_2 =& \frac{d}{4}\left\{f(x_i) + f(x_{i+\frac{1}{2}})\right\} + \frac{d}{4}\left\{f(x_{i+\frac{1}{2}}) + f(x_{i+1})\right\} \\ =& \frac{d}{4}\left\{f(x_i) + 2f(x_{i+\frac{1}{2}}) + f(x_{i+1})\right\} \\ =& \frac{1}{2}T_1 + \frac{d}{2}f(x_{i+\frac{1}{2}}) \end{aligned}$$

于是 $n+1$ 个节点的梯形公式结果 T_n 与 $2n+1$ 个节点的 T_{2n} 的关系为

$$T_{2n} = \frac{1}{2}T_n + \frac{d}{2}\sum_{i=0}^{n-1} f(x_{i+\frac{1}{2}}). \tag{4.49}$$

只需要计算新增的 n 个节点上的函数值。实际计算积分时预先确定一个误差限 ε，当 $|T_{2n} - T_n| < \varepsilon$ 时停止增加节点, 以 T_{2n} 为积分近似值。注意，如果被积函数在很大范围基本不变，比如等于零，这样的停止规则可能给出完全错误的积分计算结果。

R 的 `integrate(f, lower, upper)` 函数可以计算有限区间或无穷区间定积分，如果要指定无穷区间，取 `lower` 为`-Inf` 或取 `upper` 为Inf。R 的 MASS 包中的 `area` 函数可以计算有限区间上的定积分，结果比 `integrate` 更为可靠。

4.3.3 多维数值积分

多元函数 $f(x_1, x_2, \ldots, x_d)$ 的积分比一元函数积分困难得多，这时很难再取多元的插值多项式然后对插值函数积分，计算量也要比一维积分大得多。

对定义在矩形 $[a,b]\times[c,d]$ 上的二元函数 $f(x,y)$，为计算二重积分

$$I = \int_a^b \int_c^d f(x,y)\, dy\, dx$$

可以用如下的中点法则近似：

$$M_{n,m} = \frac{b-a}{n}\frac{d-c}{m}\sum_{i=0}^{n-1}\sum_{j=0}^{m-1} f\left(a+\left(i+\frac{1}{2}\right)h,\ c+\left(j+\frac{1}{2}\right)g\right), \tag{4.50}$$
$$\left(h=\frac{b-a}{n},\ g=\frac{d-c}{m}\right).$$

对于 R^d 空间中超立方体上的积分也可以类似计算。

高维空间的积分需要大量的计算，计算量可能会大到不适用。比如，如果一维需要 n 个节点，计算 n 次被积函数值，那么 d 维积分就需要计算 $N = n^d = e^{d\ln n}$ 次被积函数值，计算量随着 d 的增长而指数增长。随机模拟积分 (见 §3.2) 的精度为 $O_p(N^{-1/2})$，受维数 d 的影响较小，但高维时也需要计算很多点上的函数值才能相对准确地计算积分。

拟随机积分是另外一种计算积分的方法，它结合了数值积分和随机模拟积分的优点，在特殊选取的格子点上计算函数值并计算积分，这样的布点位置称为拟随机数 (quasi-random)。我国数学家方开泰和王元[17] 开创了基于数论并在试验设计、高维积分有良好应用效果的“均匀设计” 学科，利用他们提出的拟随机数布点可以进行高维积分计算。均匀设计布点有“均匀分散” 和 “整齐可比” 两个特点，使布点既具有代表性，又是有规律的。随机模拟积

分的随机误差界为 $O(\sqrt{N^{-1}\log\log N})$，而拟随机积分的误差可以达到 $O(N^{-1}(\log N)^d)$ (见 Monahan(2001)[31] §10.6, Lemieux(2009)[26])。

4.3.4 数值微分

在没有函数微分的解析表达式时，可以用数值方法由函数值近似计算函数的微分值。按照微分定义，当 $f(x)$ 在点 x 处可微时，对很小的 h 有

$$f'(x) = \frac{f(x+h)-f(x)}{h} + O(h) \tag{4.51}$$

$$= \frac{f(x)-f(x-h)}{h} + O(h) \tag{4.52}$$

$$= \frac{f(x+h)-f(x-h)}{2h} + O(h^2) \tag{4.53}$$

这三种近似分别称为数值微分的**向前差商**、**向后差商**和**中心差商公式**, 其中中心差商公式有更好的精度。h 选取应该很小，使得继续减小 h 时按差商公式计算的近似导数值不再变化或变化量小于给定误差界限。但是，h 也不能过分小，因为计算 $f(x)$ 值只能到一定精度，当 h 太小时 $f(x+h)$ 与 $f(x)$ 的差别已经不是由 h 引起的而是由 $f(x)$ 计算误差引起的。如果 $f(x)$ 的计算精度可以达到机器单位 U 量级，取 h 的一个经验法则是取 $h/|x|$ 为 $U^{1/2}$ 左右, 对中心差商公式取 $h/|x|$ 为 $U^{1/3}$ 左右 (见 Monahan(2001)[31] §8.6)。多元函数的梯度可以类似地计算。

可以类似地计算高阶微分，如

$$f''(x) = \frac{[f(x+h)-f(x)]-[f(x)-f(x-h)]}{h^2} + O(h), \tag{4.54}$$

公式分子有意地写成了两个差分的差分，以避免有效位数损失。

公式(4.51)–(4.53)是用割线斜率近似切线斜率，相当于用两点的函数值作线性插值后用插值函数的微分近似原始函数微分。一般地，对函数 $f(x)$ 可以计算插值多项式 $P_n(X)$，用 $P_n'(x)$ 近似 $f'(x)$。这种计算微分的方法称为 **插值型求导公式**。设多项式 $P_{n-1}(x)$ 满足

$$P_{n-1}(x_i) = f(x_i),\ i = 1, 2, \ldots, n$$

由定理4.2.2, 插值近似的余项为

$$f(x) - P_{n-1}(x) = \frac{f^{(n)}(\xi)}{n!} g(x) \tag{4.55}$$

其中 $g(x) = \prod_{i=1}^{n}(x - x_i)$, ξ 依赖于 x 的位置。插值型求导公式的余项为

$$f'(x) - P_{n-1}'(x) = \frac{f^{(n)}(\xi)}{n!} \cdot g'(x) + \frac{g(x)}{n!} \cdot \frac{d}{dx}\frac{f^{(n)}(\xi)}{n!} \tag{4.56}$$

因为(4.56)中的第二项中 ξ 与 x 的函数关系未知，对任意 x 处用 $P'_{n-1}(x)$ 近似 $f'(x)$ 可能有很大误差。如果仅在节点 $x_i, i=1,2,\ldots,n$ 处计算 $P'_{n-1}(x)$，则有

$$f'(x_i)=P'_n(x_i)+\frac{f^{(n)}(\xi)}{n!}g'(x_i),\ i=1,2,\ldots,n. \tag{4.57}$$

根据(4.57)，我们导出对等距节点 $(x_i,z_i)(\ i=1,2,\ldots,n)$ 计算数值微分的公式。设 $z_i=f(x_i)$，$x_i-x_{i-1}=h$。

用线性插值近似求导的公式为

$$f'(x_i)=\frac{z_{i+1}-z_i}{h}-\frac{h}{2}f''(\xi_1),\ i=1,2,\ldots,n-1\ (\xi_1\in(x_i,x_{i+1}) \tag{4.58}$$

或

$$f'(x_i)=\frac{z_i-z_{i-1}}{h}+\frac{h}{2}f''(\xi_2),\ i=2,3,\ldots,n\ (\xi_1\in(x_{i-1},x_i) \tag{4.59}$$

用二次多项式插值近似求导的公式为

$$\begin{aligned}f'(x_i)=&\frac{-3z_i+4z_{i+1}-z_{i+2}}{2h}+\frac{h^2}{3}f^{(3)}(\xi_1),\\ &i=1,2,\ldots,n-2,\ \xi_1\in(x_i,x_{i+2})\end{aligned} \tag{4.60}$$

或

$$\begin{aligned}f'(x_i)=&\frac{-z_{i-1}+z_{i+1}}{2h}-\frac{h^2}{6}f^{(3)}(\xi_2),\\ &i=2,3,\ldots,n-1,\ \xi_2\in(x_{i-1},x_{i+1})\end{aligned} \tag{4.61}$$

或

$$\begin{aligned}f'(x_i)=&\frac{z_{i-2}-4z_{i-1}+3z_i}{2h}+\frac{h^2}{3}f^{(3)}(\xi_3),\\ &i=3,4,\ldots,n,\ \xi_3\in(x_{i-2},x_i).\end{aligned} \tag{4.62}$$

用二次多项式插值近似求二阶导数的公式为

$$\begin{aligned}f''(x_i)=&\frac{z_{i-1}-2z_i+z_{i+1}}{h^2}-\frac{h^2}{12}f^{(4)}(\xi_1),\\ &i=2,3,\ldots,n-1,\ \xi_1\in(x_{i-1},x_{i+1}).\end{aligned} \tag{4.63}$$

用多项式插值方法近似计算函数微分，在一般的 x 值处计算的误差可能很大。可以采用样条函数来进行插值，用样条函数的导数来计算被插值函数的导数。

习题四

1. 设计 R 程序，用向量保存多项式的系数，设计多项式加减法、乘法、除法的程序（除法结果包括商和余式）。

2. 把 Legendre 多项式推广到 $L^2[a,b]$ 中, 给出递推公式。

3. 设计计算 $L^2[a,b]$ 上的 Legendre 多项式 P_n 系数的算法并编写 R 程序。

4. 设计计算 Laguerre 多项式 L_n 系数的算法并编写 R 程序。

5. 用 Legendre 正交多项式找到标准正态分布函数 $\Phi(x), x \in [0,3]$ 的绝对误差在 10^{-6} 以下的近似公式并用 R 程序验证。可以使用 R 中的 `pnorm` 函数作为已知的精确值。

6. 用 Laguerre 正交多项式方法找到标准正态分布函数 $\Phi(x), x \in [3,\infty)$ 的绝对误差在 10^{-6} 以下的近似公式并用 R 程序验证。

7. 证明连分式计算的正向算法。

8. 编写用连分式公式(4.15)计算标准正态分布函数 $\Phi(x)$ 的 R 程序，要求绝对误差控制在 10^{-10} 以下。

9. 编写 R 程序，对输入的 n 个点在给定的自变量 x 处进行抛物线插值，x 为向量。考虑 n 个点的自变量等距和不等距两种情况。

10. 证明复合辛普森积分公式(4.38)。

11. 编写变步长梯形法则积分算法和相应 R 程序。

12. 把变步长梯形法则积分算法推广到 $[0,\infty)$ 上的定积分：设已用 n 等分计算 $[0,b]$ 上的积分，下一步增加原有 n 个区间的中点作为节点，并在 $(b,2b]$ 增加 $2n$ 个节点，计算 $[0,2b]$ 上的定积分，直到结果变化小于给定的误差限 ε。用 $\int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \frac{1}{2}$ 验证。

13. 在贝叶斯分析中先验分布为 $\pi(\theta) = 6\theta(1-\theta)$, 似然函数为 $L(\theta) \propto \theta^{15}/(-\log(1-\theta))^{10}$, 后验密度为（差一个比例系数）

$$p^*(\theta) \propto \theta^{16}(1-\theta)(-\log(1-\theta))^{10}.$$

分别对 $h_0(\theta) = 1$, $h_1(\theta) = \theta$, $h_2(\theta) = \theta^2$ 计算数值积分 $\int_0^1 h_i(\theta)p^*(\theta)d\theta$, 分别用中点法则、复合梯形法则和复合辛普森法则计算，比较达到小数点后 1 位的精度所需要的节点个数。

14. 令 $\boldsymbol{x} = (x_1, x_2, x_3, x_4)^T$, $f(\boldsymbol{x}) = [\log(x_1 x_2)/(x_3 x_4)]^2$, $x_i > 0, i = 1, 2, 3, 4$。求 $f(\boldsymbol{x})$ 的梯度 $\nabla f(\boldsymbol{x})$ 的表达式，并编写 R 程序用数值微分方法估计 $\nabla f(\boldsymbol{x})$，比较两者的结果。

第五章　矩阵计算

5.1　介绍

统计模型和统计计算中广泛使用矩阵运算和线性方程组求解。例如，如下的线性模型

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.1}$$

中，$\boldsymbol{y}$ 为 $n \times 1$ 向量, X 为 $n \times p$ 矩阵, 一般第一列元素全是 1, 代表截距项；$\boldsymbol{\beta}$ 为 $p \times 1$ 未知参数向量; $\boldsymbol{\varepsilon}$ 为 $n \times 1$ 随机误差向量, $\boldsymbol{\varepsilon}$ 的元素独立且方差为相等的 σ^2(未知)。参数 $\boldsymbol{\beta}$ 的最小二乘估计是如下正规方程的解：

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}, \tag{5.2}$$

当 X 为列满秩矩阵时 $\boldsymbol{\beta}$ 的最小二乘估计可以表示为

$$\boldsymbol{\beta} = (X^T X)^{-1} X^T\boldsymbol{y} \tag{5.3}$$

因变量 $\boldsymbol{y}$ 的拟合值（预报值）可以表示为

$$\hat{\boldsymbol{y}} = X(X^T X)^{-1} X^T\boldsymbol{y} = H\boldsymbol{y}, \tag{5.4}$$

其中 $H = X(X^T X)^{-1} X^T$ 是对称幂等矩阵。

再比如，在时间序列分析问题中需要对多元序列建模，常用的一种模型是向量自回归模型，p 阶向量自回归模型可以写成

$$\boldsymbol{x}_t = \sum_{j=1}^{p} A_j \boldsymbol{x}_{t-j} + \boldsymbol{\varepsilon}_t,\ t \in \mathbb{Z} \tag{5.5}$$

其中 $\boldsymbol{x}_t$ 是 m 元随机向量，$A_1, A_2, \ldots, A_p$ 是 $m \times m$ 矩阵，$\boldsymbol{\varepsilon}_t$ 是 m 元白噪声。

可见，统计模型中广泛使用矩阵作为模型表达工具，统计计算中有大量的矩阵计算。我们需要研究稳定、高效的矩阵计算方法。例如，按照(5.3)计算 $\hat{\boldsymbol{\beta}}$ 从理论上很简单，但需要计

算逆矩阵，实际计算量比较大；如果把(5.2)看成是有多个列向量作为等式右边的线性方程组来求解，则可以找到各种快速且高精度的计算方法。

在本书中我们用黑体小写字母表示向量，且缺省为列向量，如 $\boldsymbol{a}, \boldsymbol{v}$，用 a_i 表示 $\boldsymbol{a}$ 的第 i 元素。用 $\mathbb{R}^n$ 表示所有 n 维实值向量组成的 n 维欧式空间, 用 $(\boldsymbol{a}, \boldsymbol{b})$ 表示 $\boldsymbol{a}^T\boldsymbol{b}$，称为 $\boldsymbol{a}, \boldsymbol{b}$ 的内积，记 $\|\boldsymbol{a}\| = (\boldsymbol{a}, \boldsymbol{a})^{1/2}$，称为 $\boldsymbol{a}$ 的欧式模。用大写字母表示矩阵，如 A, M，用 a_{ij} 表示 A 的第 i 行第 j 列元素，用 $\boldsymbol{a}_{\cdot j}$ 表示 A 的第 j 列组成的列向量，用 $\boldsymbol{a}_{i\cdot}$ 表示 A 的第 i 行组成的行向量。用 A^T 表示 A 的转置, $\det(A)$ 表示 A 的行列式。另外，一些特殊的矩阵定义如下：

- 设 $\boldsymbol{e}_i$ 为 n 维列向量，如果其第 i 个元素为 1，其它元素为 0，称 $\boldsymbol{e}_i$ 为 n 维**单位向量**。
- 记 $\mathbf{1}$ 为元素都是 1 的列向量。
- 用 I_n 表示 n 阶单位阵，用 I 表示单位阵。
- 若矩阵 A 的元素满足 $a_{ij} = 0, \forall i < j$，称 A 为**上三角矩阵**。
- 若矩阵 A 的元素满足 $a_{ij} = 0, \forall i > j$，称 A 为**上三角矩阵**。
- 若矩阵 A 的元素满足 $a_{ij} = 0, \forall i \neq j$，称 A 为**对角矩阵**。
- 若矩阵 A 的元素满足 $a_{ij} = 0, \forall |i-j| > 1$，称 A 为**三对角矩阵**。
- 若实方阵 A 满足 $A^TA = I$，则称 A 为**正交阵**。
- 若 A 为 n 阶实对称矩阵，对任意 n 维非零实数向量 $\boldsymbol{x}$ 有 $\boldsymbol{x}^TA\boldsymbol{x} > 0$，称 A 为**正定阵**。
- 若 A 为 n 阶实数对称矩阵，对任意 n 维实数向量 $\boldsymbol{x}$ 有 $\boldsymbol{x}^TA\boldsymbol{x} \geq 0$，称 A 为**非负定阵**或**半正定阵**。
- 若 $P(i,j)$ 是把 I_n 的第 i 行和第 j 行交换位置后得到的 n 阶方阵，称 $P(i,j)$ 为**基本置换阵**。$P(i,j)A$ 把 A 的第 i、j 两行互换, $AP(i,j)$ 把 A 的第 i、j 两列互换。$P(i,j)P(i,j) = I_n$。
- 设 $\boldsymbol{\pi} = (k_1, k_2, \ldots, k_n)^T$ 是由 $(1, 2, \ldots, n)$ 的一个排列组成的 n 维向量，方阵 $P = (\boldsymbol{e}_{k_1}, \boldsymbol{e}_{k_2}, \ldots, \boldsymbol{e}_{k_n})$, 称 P 为**置换阵**。P 是一个正交矩阵，对任意 $m \times n$ 矩阵 A, AP 是把 A 的各列按照 $k_1, k_2, \ldots, k_n$ 次序重新排列得到的矩阵。
- 设 A 为 $n \times m$ 矩阵 $(n \geq m)$，称 $\mu(A) \triangleq \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = A\boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^m\}$ 为由 A 的列向量张成的线性子空间。

- 设 n 阶实对称矩阵 P 满足 $P^2 = P$，称 P 为**对称幂等矩阵**，P 是 $\mathbb{R}^n$ 到 $\mu(P)$ 的 (**正交**) **投影矩阵**，对任意 $\boldsymbol{x} \in \mathbb{R}^n$, $P\boldsymbol{x}$ 与 $\boldsymbol{x} - P\boldsymbol{x} = (I - P)\boldsymbol{x}$ 正交。若 A 为 $n \times m$ 列满秩矩阵, 则 $P = A(A^TA)^{-1}A^T$ 是 $\mathbb{R}^n$ 到 $\mu(A)$ 的正交投影矩阵。

这里列出概率论中关于随机向量和随机矩阵的几个基本公式。设 $\boldsymbol{X}$ 为 n 元随机向量，$\boldsymbol{M}$ 为 $m \times n$ 随机矩阵，A, B 为普通非随机的实数矩阵。设 $\mathrm{Var}(\boldsymbol{X})$ 表示 $\boldsymbol{X}$ 的协方差阵。有如下常用公式:

$$\begin{aligned} E(A\boldsymbol{M}B) =& AE(\boldsymbol{M})B \\ \mathrm{Var}(\boldsymbol{X}) =& E\left[(\boldsymbol{X} - E\boldsymbol{X})(\boldsymbol{X} - E\boldsymbol{X})^T\right] \\ \mathrm{Var}(A\boldsymbol{X}) =& A\mathrm{Var}(\boldsymbol{X})A^T \end{aligned}$$

有许多编程语言有成熟的矩阵计算软件包，比如 FORTRAN 和 C 语言的 LAPACK 程序包[12]、IMSL[23] 程序包。在常用统计软件系统一般内建了高等矩阵计算功能，比如 R 软件 (见 §1.2)、SAS 软件中的 IML 模块、MATLAB 软件，等等。我们可能不需要再去自己编写矩阵乘法、解线性方程组这些基础的计算程序，但是还是要了解这里面涉及的算法，这样遇到高强度、高维数等复杂情形下的矩阵计算问题才能给出稳定、高效的解决方案。对于反复使用的矩阵运算，1% 的速度提升也是难能可贵的；对于高阶矩阵，应该尽可能少产生中间结果矩阵，尽可能把输入和输出保存在同一存储位置。

例 5.1.1. 设 A 为 $n \times n$ 矩阵，$\boldsymbol{x}$ 为 n 维列向量，计算矩阵乘法 $A\boldsymbol{x}$ 需要 n^2 次乘法和 n^2 次加法。如果 A 有特殊结构 $A = I_n + \boldsymbol{u}\boldsymbol{v}^T$，其中 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 是 n 维列向量，则

$$A\boldsymbol{x} = \boldsymbol{x} + (\boldsymbol{v}^T\boldsymbol{x})\boldsymbol{u}$$

只需要 $2n$ 次乘法和 $2n$ 次加法，并且矩阵 A 也不需要保存 n^2 个元素，而只需 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的 $2n$ 个元素。若 A 是上三角矩阵或下三角矩阵，则 $A\boldsymbol{x}$ 只需要 $\frac{1}{2}n(n+1)$ 次乘法和加法，计算量比一般的 A 减少一半。

在 R 语言中，用 `matrix` 函数定义一个矩阵，用 `cbind` 和 `rbind` 进行横向和纵向合并，用 `t(A)` 表示 A 的转置，用`A %*% B` 表示矩阵 A 和 B 相乘。

5.2 线性方程组求解

统计计算和其它科学工程计算中很多的问题会转化为线性方程组求解问题。本节讨论稳定、高效的线性方程组求解方法。

5.2.1 三角形线性方程组求解

我们熟知的解线性方程组的一种方法是消元法，用线性变化把增广矩阵化为阶梯形然后用回代法求解。我们先给出系数矩阵为三角形矩阵的线性方程组解法。

设有如下的三角形 n 阶线性方程组

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1,n-1}x_{n-1} + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \cdots + a_{2,n-1}x_{n-1} + a_{2n}x_n &= b_2 \\ \ddots \qquad\qquad & \vdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n &= b_{n-1} \\ a_{nn}x_n &= b_n \end{aligned}$$

设其系数矩阵满秩（当且仅当 $a_{11}a_{22}\ldots a_{nn} \neq 0$），则求解过程可以写成：

$$\begin{aligned} x_n &= b_n/a_{nn} \\ x_{n-1} &= \left(b_{n-1} - a_{n-1,n}x_n\right)/a_{n-1,n-1} \\ &\vdots \\ x_2 &= \left(b_2 - a_{23}x_3 - \cdots - a_{2n}x_n\right)/a_{22} \\ x_1 &= \left(b_1 - a_{12}x_2 - \cdots - a_{1,n}x_n\right)/a_{11} \end{aligned}$$

这种求解过程叫做**回代法**。需要的话可以把解出的 $\boldsymbol{x}$ 元素保存在 $\boldsymbol{b}$ 原来的存储空间中。

如果系数矩阵是下三角的，也可以类似求解。

5.2.2 高斯消元法和 LU 分解

高斯消元法是众所周知的线性方程组解法，配合主元元素选取可以得到稳定的解。

例 5.2.1. 考虑线性方程组

$$\left\{\begin{array}{rcrcrcrcr} 3x_1 & + & x_2 & + & 2x_3 & + & x_4 & = & 5 \\ 6x_1 & + & 4x_2 & + & 7x_3 & + & 11x_4 & = & 5 \\ 15x_1 & + & 11x_2 & + & 18x_3 & + & 34x_4 & = & 6 \\ 18x_1 & + & 16x_2 & + & 25x_3 & + & 56x_4 & = & -4 \end{array}\right. \tag{5.6}$$

只要把方程组变成上三角形就可以用 §5.2.1描述的方法求解。记

$$A=\begin{pmatrix}3&1&2&1\\6&4&7&11\\15&11&18&34\\18&16&25&56\end{pmatrix}\quad \boldsymbol{b}=\begin{pmatrix}5\\5\\6\\-4\end{pmatrix}\quad \overline{A}=(A\,|\,\boldsymbol{b})\quad \boldsymbol{x}=\begin{pmatrix}x_1\\x_2\\x_3\\x_4\end{pmatrix}$$

令 $\boldsymbol{m}^{(1)}=(0,2,5,6)^T$，把第一个方程乘以 $-m_2^{(1)}=-2$ 加到第二个方程上，可以消去第二个方程中 x_1 的系数。类似地，把第一个方程乘以 $-m_3^{(1)}=-5$ 加到第三个方程上，把第一个方程乘以 $-m_4^{(1)}=-6$ 加到第四个方程上，可以消去第三、第四方程中 x_1 的系数。方程组变成

$$\begin{cases}3x_1+x_2+2x_3+x_4=5\\ 2x_2+3x_3+9x_4=-5\\ 6x_2+8x_3+29x_4=-19\\ 10x_2+13x_3+50x_4=-34\end{cases}\tag{5.7}$$

记 $\boldsymbol{m}^{(2)}=(0,0,3,5)^T$，把第二个方程乘以 $-m_3^{(2)}=-3$ 加到第三个方程上，可以消去第三个方程中 x_2 的系数；把第二个方程乘以 $-m_4^{(2)}=-5$ 加到第四个方程上，可以消去第四个方程中 x_2 的系数。方程组变成

$$\begin{cases}3x_1+x_2+2x_3+x_4=5\\ 2x_2+3x_3+9x_4=-5\\ -x_3+2x_4=-4\\ -2x_3+5x_4=-9\end{cases}\tag{5.8}$$

记 $\boldsymbol{m}^{(3)}=(0,0,0,2)^T$, 把第三个方程乘以 $-m_4^{(3)}=-2$ 加到第四个方程上，可以消去第四个方程中 x_3 的系数。方程组变成

$$\begin{cases}3x_1+x_2+2x_3+x_4=5\\ 2x_2+3x_3+9x_4=-5\\ -x_3+2x_4=-4\\ x_4=-1\end{cases}\tag{5.9}$$

用回代法可以解出 $\boldsymbol{x}=(1,-1,2,-1)^T$。

考虑例5.2.1把系数矩阵变成上三角形的过程。第 j 步用初等变换把系数矩阵第 j 列的主对角线下方的元素都变成零。设 $A^{(0)}=A$，A 为 $n\times n$ 系数矩阵，第 j 步把矩阵 $A^{(j-1)}$ 变成矩阵 $A^{(j)}$，实际是左乘了一个初等变换矩阵 $M^{(j)}$:

$$A^{(j)}=M^{(j)}A^{(j-1)},\ j=1,2,\ldots,n-1\tag{5.10}$$

其中 $M^{(j)}$ 是一个与 I_n 仅在第 j 列不同的方阵，其第 j 列为

$$m_{ij}^{(j)} = \begin{cases} 0, & \text{当} i < j \\ 1, & \text{当} i = j \\ -a_{ij}^{(j-1)}/a_{jj}^{(j-1)}, & \text{当} i > j \end{cases} \tag{5.11}$$

定义 n 维向量 $\boldsymbol{m}^{(j)}$ 为

$$m_{i}^{(j)} = \begin{cases} 0, & \text{当} i \le j \\ a_{ij}^{(j-1)}/a_{jj}^{(j-1)}, & \text{当} i > j \end{cases} \tag{5.12}$$

则初等变换矩阵 $M^{(j)}$ 可以表示为

$$M^{(j)} = I_n - \boldsymbol{m}^{(j)}\boldsymbol{e}_j^T \tag{5.13}$$

经过 $n-1$ 步消去变换后，得到的上三角形系数矩阵为

$$A^{(n-1)} = M^{(n-1)} \dots M^{(2)} M^{(1)} A \tag{5.14}$$

如果每一步中的分母 $a_{jj}^{(j-1)}$ 都不等于零，则上述步骤可以把方程组的系数矩阵变成上三角形, 对线性方程组右边的 $\boldsymbol{b}$ 也作相同的变换就得到三角形线性方程组，可以用回代法求解。但是，$a_{jj}^{(j-1)} = 0$ 的情况是存在的，例如

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \tag{5.15}$$

就无法用例5.2.1的方法化为上三角形。

这种情况怎么办？当第 j 步的 $a_{jj}^{(j-1)} = 0$ 时，我们可以看第 $j+1, j+2, \dots, n$ 个方程中 x_j 的系数是否非零，如果第 s_j 个方程的 x_j 系数不等于零，可以把第 j 个方程与第 s_j 个方程互换，然后继续消元。把第 j 个方程与第 s_j 个方程互换的操作相当于对系数矩阵左乘简单置换阵 $P(j, s_j)$。那么，是不是第 $j+1, j+2, \dots, n$ 个方程中只要 x_j 的系数非零就可以与第 j 个方程互换？要注意的是，第 j 个方程与第 s_j 个方程互换后原来的 $a_{s_j j}^{(j-1)}$ 就变成了第 j 列的第 j 行元素，要作为消去的分母 (称为**主元**)；而数值计算中 “非零” 的判断是很不可靠的，应该等于零的值在数值计算中很可能因为舍入误差变成非零。所以，在轮到对第 j 列消元时，不论 $a_{jj}^{(j-1)}$ 是否等于零，都要从系数矩阵第 j 列的第 $j, j+1, \dots, n$ 行元素中找到绝对值最大的一个，假设此元素在第 j 列的第 s_j 行，就把第 s_j 行与第 j 行互换，然后进行消去。这

种解方程的方法叫做**列主元的高斯消元法**。得到三角形矩阵及相应方程组的过程为:

$$A^{(n-1)} = M^{(n-1)}P(n-1,s_{n-1})\dots M^{(2)}P(2,s_2)M^{(1)}P(1,s_1)A \tag{5.16}$$

$$A^{(n-1)}\boldsymbol{x} = M^{(n-1)}P(n-1,s_{n-1})\dots M^{(2)}P(2,s_2)M^{(1)}P(1,s_1)\boldsymbol{b} \tag{5.17}$$

注意，其中的 $M^{(j)}$ 是基于行置换后的 $P(j,s_j)A^{(j-1)}$ 计算的。实际求解时，通行的算法是把每一步的 $A^{(j)}$ 保存在 A 的存储空间，但已经消去变成零的元素不保存，代之以将 $\boldsymbol{m}^{(j)}$ 的第 $j+1,j+2,\dots,n$ 号元素存放在 $A^{(j)}$ 第 j 列的最后 $n-j$ 个元素中 (这 $n-j$ 本来被消去变成了零), 解 $\boldsymbol{x}$ 的元素可以存放在 $\boldsymbol{b}$ 的存储空间中。要注意的是，在第 $j+1,j+2,\dots$ 步时的行置换会打乱这样保存的 $\boldsymbol{m}^{(j)}$ 的元素次序。

可以证明 (见习题5)，按照以上所述的步骤，把原始矩阵 A 做了如下 LU 分解:

$$PA = LU \tag{5.18}$$

其中 P 是一个置换矩阵:

$$P = P(n-1,s_{n-1})\dots P(2,s_2)P(1,s_1), \tag{5.19}$$

只要保存向量 $(s_1,s_2,\dots,s_{n-1})$ 就可以恢复矩阵 P。U 是一个上三角矩阵，其上三角元素保存在原来输入的 A 存储空间的上三角部分，为列主元法消元最后得到的 $A^{(n-1)}$ 矩阵。L 是一个单位下三角矩阵（主对角线元素都等于 1 的下三角矩阵），它严格下三角部分（不包括主对角线在内的下三角部分）的元素保存在原来输入的 A 的存储空间的严格下三角部分，第 j 列的最后的 $n-j$ 个元素保存了被第 $j+1,j+2,\dots,n-1$ 次行置换打乱次序的 $\boldsymbol{m}^{(j)}$, 记为 $\boldsymbol{m}_*^{(j)}$:

$$\boldsymbol{m}_*^{(j)} = P(n-1,s_{n-1})\dots P(j+1,s_{j+1})\boldsymbol{m}^{(j)}, \tag{5.20}$$

所以 L 可以表示为

$$L = I_n + \boldsymbol{m}_*^{(1)}\boldsymbol{e}_1^T + \dots + \boldsymbol{m}_*^{(n-1)}\boldsymbol{e}_{n-1}^T. \tag{5.21}$$

公式(5.18)称为矩阵 A 的三角分解或 LU 分解。

在得到 LU 分解(5.18)后，对任何一个常数向量 $\boldsymbol{b}$，要求解 $A\boldsymbol{x}=\boldsymbol{b}$，可以化为 $PA\boldsymbol{x}=P\boldsymbol{b}$，注意左乘矩阵 P 只是进行了 $n-1$ 次行置换。由 $PA=LU$ 得 $LU\boldsymbol{x}=(P\boldsymbol{b})$, 令 $\boldsymbol{y}=U\boldsymbol{x}$，先用回代法解下三角形方程组 $L\boldsymbol{y}=P\boldsymbol{b}$，再用回代法解上三角形方程组 $U\boldsymbol{x}=\boldsymbol{y}$ 就可以得到方程组的解。

求解线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 时，为什么不先求逆矩阵 A^{-1} 再求 $\boldsymbol{x}=A^{-1}\boldsymbol{b}$? 仔细分析上述算法的运算次数可以发现，得到 LU 分解(5.18)只需 $\frac{2}{3}n^3+O(n^2)$ 次浮点运算（加减乘除），

额外地再用两遍回代法对一个 $\boldsymbol{b}$ 求解 $\boldsymbol{x}$ 仅需 $2n^2$ 次浮点运算，但是在得到(5.18)后如果要求 A^{-1} 则需要额外的 $\frac{2}{3}n^3$ 次浮点运算 (见习题4)，即求 A^{-1} 需要 $\frac{4}{3}n^3+O(n^2)$ 次浮点运算。直接用消元法求 A^{-1} 也需要 $\frac{4}{3}n^3+O(n^2)$ 次浮点运算。得到 A^{-1} 后计算 $\boldsymbol{x}=A^{-1}\boldsymbol{b}$ 需要 $2n^2$ 次浮点运算，与两遍回代法求 $\boldsymbol{x}$ 计算量相同, 但多出了求逆的 $\frac{2}{3}n^3$ 次浮点运算。所以如果仅需要解线性方程组的话不需要计算逆矩阵。

得到 LU 分解(5.18)后可以给出 A 的行列式：

$$\det(P)\det(A)=\det(PA)=\det(LU)=\det(U) \tag{5.22}$$

$\det(U)$ 等于 U 的主对角线元素乘积，$\det(P)$ 等于 1 或 -1，当 P 代表偶数次行置换时 $\det(P)=1$，否则 $\det(P)=-1$。注意 P 是 $n-1$ 个行置换的乘积，当 $s_j=j$ 时 $P(j,s_j)=I_n$，$\det(P(j,s_j))=1$；当 $s_j\neq j$ 时 $\det(P(j,s_j))=-1$。所以，如果向量 $(s_1,s_2,\ldots,s_{n-1})$ 与向量 $(1,2,\ldots,n-1)$ 不相等的元素个数是偶数，则 $\det(P)=1$，否则 $\det(P)=-1$。

在 R 语言中，用 `solve(A, b)` 求解 $A\boldsymbol{x}=\boldsymbol{b}$, 用 `solve(A)` 求 A 的逆矩阵，用 Matrix 包的 `lu` 函数求 LU 分解。

5.2.3 Cholesky 分解

统计计算中比一般矩阵更常用到的是非负定矩阵和正定矩阵。比如，随机向量 $\boldsymbol{X}$ 的协方差阵 $\Sigma=\mathrm{Var}(\boldsymbol{X})$ 是非负定的，如果满秩，就是正定的。回归分析中正规方程(5.2)中的叉积阵 X^TX 是非负定的，如果 X 列满秩则 X^TX 是正定的。如果方程 $A\boldsymbol{x}=\boldsymbol{b}$ 的系数矩阵 A 是正定矩阵，虽然还可以用高斯消元法或 LU 分解来求解，但这样不能利用 A 的特殊结构。正定矩阵 A 可以进行所谓 Cholesky 分解:

$$A=LL^T \tag{5.23}$$

其中 L 是一个主对角线元素都取正值的下三角阵。

先承认(5.23)的存在性, 设法求 L。设 L 的元素为 l_{ij}, $i\geq j$, 则

$$A=\begin{pmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ \vdots & \vdots & \ddots & \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix}\begin{pmatrix} l_{11} & l_{21} & \cdots & l_{n1} \\ & l_{22} & \cdots & l_{n2} \\ & & \ddots & \vdots \\ & & & l_{nn} \end{pmatrix} \tag{5.24}$$

显然 $a_{11}=l_{11}^2$。用 $A^{[k]}$ 表示 A 的前 k 行和前 k 列组成的子矩阵，$L^{[k]}$ 表示 L 的前 k 行和前 k 列组成的子矩阵，易见 $A^{[k]}=L^{[k]}(L^{[k]})^T$。归纳地，设 $L^{[k-1]}$ 已经求得，要求 L 的第 k 行,

记 $(\boldsymbol{l}^{[k]})^T=(l_{k1},l_{k2},\ldots,l_{k,k-1})$, $\boldsymbol{a}^{[k]}=(a_{1,k},a_{2,k},\ldots,a_{k-1,k})^T$, 由 $A^{[k]}=L^{[k]}(L^{[k]})^T$ 有

$$\begin{aligned}A^{[k]}=&\begin{pmatrix}A^{[k-1]} & \boldsymbol{a}^{[k]}\\ (\boldsymbol{a}^{[k]})^T & a_{kk}\end{pmatrix}\\ =&L^{[k]}(L^{[k]})^T=\begin{pmatrix}L^{[k-1]} & \boldsymbol{0}\\ (\boldsymbol{l}^{[k]})^T & l_{kk}\end{pmatrix}\begin{pmatrix}(L^{[k-1]})^T & \boldsymbol{l}^{[k]}\\ \boldsymbol{0} & l_{kk}\end{pmatrix}\end{aligned}\tag{5.25}$$

得方程

$$L^{[k-1]}\boldsymbol{l}^{[k]}=\boldsymbol{a}^{[k]}\tag{5.26}$$

$$l_{kk}^2=a_{kk}-(\boldsymbol{l}^{[k]})^T\boldsymbol{l}^{[k]}\tag{5.27}$$

只要用回代法解下三角方程组(5.26)得到 $\boldsymbol{l}^{[k]}$, 然后开平方根得到 $l_{kk}=\left(a_{kk}-(\boldsymbol{l}^{[k]})^T\boldsymbol{l}^{[k]}\right)^{1/2}$。

例 5.2.2. 矩阵

$$A=\begin{pmatrix}4 & 2 & 0 & 2\\ 2 & 10 & 12 & 1\\ 0 & 12 & 17 & 2\\ 2 & 1 & 2 & 9\end{pmatrix}\tag{5.28}$$

是正定阵，来求它的 Cholesky 分解。

k=1: $l_{11}=\sqrt{a_{11}}=\sqrt{4}=2.$

k=2: 方程 $L^{[1]}\boldsymbol{l}^{[2]}=\boldsymbol{a}^{[2]}$ 即 $l_{11}l_{21}=a_{21}$(注意 $a_{12}=a_{21}$), 所以 $2l_{21}=2$, $l_{21}=1$; 于是用(5.27)得 $l_{22}=\sqrt{a_{22}-l_{21}^2}=\sqrt{10-1^2}=3$;

k=3: 方程 $L^{[2]}\boldsymbol{l}^{[3]}=\boldsymbol{a}^{[3]}$ 即

$$\begin{pmatrix}2 & 0\\ 1 & 3\end{pmatrix}\begin{pmatrix}l_{31}\\ l_{32}\end{pmatrix}=\begin{pmatrix}0\\ 12\end{pmatrix}$$

解得 $l_{31}=0$, $l_{32}=4$, 再求出 $l_{33}=\sqrt{a_{33}-(l_{31}^2+l_{32}^2)}=\sqrt{17-(0^2+4^2)}=1$。

k=4: 方程 $L^{[3]}\boldsymbol{l}^{[4]}=\boldsymbol{a}^{[4]}$ 即

$$\begin{pmatrix}2 & 0 & 0\\ 1 & 3 & 0\\ 0 & 4 & 1\end{pmatrix}\begin{pmatrix}l_{41}\\ l_{42}\\ l_{43}\end{pmatrix}=\begin{pmatrix}2\\ 1\\ 2\end{pmatrix}$$

解得 $l_{41}=1$, $l_{42}=0$, $l_{43}=2$, 再开平方根得到 $l_{44}=\sqrt{a_{44}-(l_{41}^2+l_{42}^2+l_{43}^2)}=2$。

于是，

$$L=\begin{pmatrix}2&&&\\1&3&&\\0&4&1&\\1&0&2&2\end{pmatrix},\quad A=LL^T$$

□

当 A 为正定阵时，各 $A^{[k]}$ 也是正定阵，取向量

$$\boldsymbol{\alpha}=\begin{pmatrix}-(A^{[k-1]})^{-1}\boldsymbol{a}^{[k]}\\1\\\mathbf{0}\end{pmatrix}\tag{5.29}$$

则(5.27)的右边是二次型 $\boldsymbol{\alpha}^TA\boldsymbol{\alpha}$，应该为正值，所以解方程(5.26)–(5.27)逐次得到的 $a_{kk}>0$，这样 $L^{[k]}$ 是满秩的，于是下一步的方程(5.26)有唯一解，先计算 $l_{11}=\sqrt{a_{11}}$ 然后对 $k=2,3,\ldots,n$ 重复解(5.26)–(5.27) 一定每步都可以进行且解是唯一的。于是，正定阵 A 的 Cholesky 分解是存在唯一的。

Cholesky 分解需要 $\frac{1}{3}n^3+O(n^2)$ 次浮点运算和 n 次开平方根，比 LU 分解的 $\frac{2}{3}n^3$ 次浮点运算次数少，开平方根所需的时间只是 n 的倍数。

编写 Cholesky 算法的程序时，如果输入了一个正定阵 A, 因为正定阵是对称的，可以只输入 A 的下三角部分，返回的 Cholesky 分解 L^TL 的下三角矩阵 L 可以保存在输入的矩阵 A 的下三角部分的存储空间中。对于对称矩阵和下三角矩阵也可以只保存其下三角部分的元素，这样按行排列的话，第 (i,j) 元素存放在第 $\frac{1}{2}i(i-1)+j$ 号位置。

统计中经常需要计算正定矩阵逆矩阵的二次型 $\boldsymbol{\alpha}^TA^{-1}\boldsymbol{\alpha}$, 就可以用 Cholesky 分解转化为

$$\boldsymbol{\alpha}^TA^{-1}\boldsymbol{\alpha}=\boldsymbol{\alpha}^T(LL^T)^{-1}\boldsymbol{\alpha}=(L^{-1}\boldsymbol{\alpha})^TL^{-1}\boldsymbol{\alpha},\tag{5.30}$$

只要用回代法解 $L\boldsymbol{x}=\boldsymbol{\alpha}$, 则有 $\boldsymbol{\alpha}^TA^{-1}\boldsymbol{\alpha}=\boldsymbol{x}^T\boldsymbol{x}$。这样计算只需要 Cholesky 分解的 $\frac{1}{3}n^3+O(n^2)$ 次运算，如果直接计算逆矩阵，则需要 $\frac{4}{3}n^3+O(n^2)$ 次运算。

在 R 软件中，用 `chol()` 函数求正定阵的 Cholesky 分解。

设 A 的 Cholesky 分解 $A=LL^T$ 中 L 的主对角线元素组成的对角阵为 $D=\mathrm{diag}(l_{11},l_{22},\ldots,l_{nn})$, 令 $\tilde{L}=LD^{-1}$, $\tilde{D}=D^2$, 则 $\tilde{L}$ 为单位下三角阵，$\tilde{D}$ 是主对角线元素为正值的对角阵，A 有如下分解:

$$A=LL^T=\tilde{L}DD\tilde{L}^T=\tilde{L}\tilde{D}\tilde{L}^T,\tag{5.31}$$

当 A 为对称正定阵时此分解存在唯一，称为矩阵 A 的 LDL 分解。

5.2.4 线性方程组求解的稳定性

设 A 为 n 阶方阵，则线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 存在唯一解的充分必要条件是 A 满秩，即 A 的列向量的任何一个非零线性组合都不等于零向量。但是，如果存在 A 的列向量的一个非零线性组合与零向量十分接近会怎样？这时，方程求解的误差很大而且不稳定（计算误差和存储误差的影响很大）。

例 5.2.3. 方程组

$$\begin{pmatrix} 3 & 1 \\ 3.0001 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} 4 \\ 4.0001 \end{pmatrix}$$

有精确解 $(1,1)^T$。对 $A,\boldsymbol{b}$ 作微小变化：

$$\begin{pmatrix} 3 & 1 \\ 2.9999 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}=\begin{pmatrix} 4 \\ 4.0002 \end{pmatrix}$$

则精确解变为 $(-2,10)$。这里, $\det(A)=-0.0001$, A 已经很接近于不满秩。

为了考察线性方程组求解的稳定性，首先回顾一些关于向量和矩阵运算的内容。设 $\boldsymbol{x}\in\mathbb{R}^n$, 定义

$$\|\boldsymbol{x}\|_p=\left(\sum_{i=1}^{n}|x_i|^p\right)^{1/p}, \tag{5.32}$$

这叫做向量 $\boldsymbol{x}$ 的 p 范数。特别地，$p=1,2,+\infty$ 时有

$$\|\boldsymbol{x}\|_1=\sum_{i=1}^{n}|x_i|, \tag{5.33}$$

$$\|\boldsymbol{x}\|_2=\sqrt{\sum_{i=1}^{n}x_i^2}=\sqrt{\boldsymbol{x}^T\boldsymbol{x}}, \tag{5.34}$$

$$\|\boldsymbol{x}\|_\infty=\max_{1\le i\le n}|x_i|. \tag{5.35}$$

向量范数作为向量空间中广义的长度，满足: $\|\boldsymbol{x}\|\ge 0$, 等号成立当且仅当 $\boldsymbol{x}=\boldsymbol{0}$; 对任意 $\lambda\in\mathbb{R}$ 有 $\|\lambda\boldsymbol{x}\|=|\lambda|\cdot\|\boldsymbol{x}\|$; 有三角不等式 $\|\boldsymbol{x}+\boldsymbol{y}\|\le\|\boldsymbol{x}\|+\|\boldsymbol{y}\|$。

给定矩阵 A, $f(\boldsymbol{x})=A\boldsymbol{x}$ 是一个多元线性函数，或称为一个线性算子。为了考察 $\boldsymbol{x}$ 的变化引起的 $A\boldsymbol{x}$ 的变化大小，引入矩阵范数 $\|A\|$。定义

$$\|A\|_p=\sup_{\|\boldsymbol{x}\|_p=1}\|A\boldsymbol{x}\|, \tag{5.36}$$

称为 A 的 p 范数。可以证明 (见习题12)，

$$\|A\|_1 = \max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|, \tag{5.37}$$

$$\|A\|_2 = \sqrt{A^TA\text{的最大特征值}}, \tag{5.38}$$

$$\|A\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|a_{ij}|. \tag{5.39}$$

定义矩阵范数后有

$$\|A\boldsymbol{x}\| \le \|A\|\cdot\|\boldsymbol{x}\|,\ \forall \boldsymbol{x}\in\mathbb{R}^n, \tag{5.40}$$

进一步

$$\|AB\boldsymbol{x}\| \le \|A\|\cdot\|B\boldsymbol{x}\| \le \|A\|\cdot\|B\|\cdot\|\boldsymbol{x}\| \tag{5.41}$$

所以 $\|AB\| \le \|A\|\cdot\|B\|$。

一般的矩阵范数定义如下。

定义 (**矩阵范数**) 若对任意 n 阶实方阵 A，都有一个实数 $\|A\|$ 与之对应，且满足

(1) 非负性: $\|A\| \ge 0$, 且等号成立当且仅当 A 为零矩阵;

(2) 齐次性: 对任意 $\lambda\in\mathbb{R}$, 有 $\|\lambda A\| = |\lambda|\,\|A\|$;

(3) 三角不等式: 对任意 n 阶实方阵 A 和 B，都有 $\|A+B\| \le \|A\| + \|B\|$;

(4) 相容性: 对任意 n 阶实方阵 A, B, 都有 $\|AB\| \le \|A\|\,\|B\|$,

则称 $\|A\|$ 为 A 的矩阵范数。

定义

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^2\right)^{1/2}, \tag{5.42}$$

则 $\|A\|_F$ 是 A 的一个矩阵范数，叫做 Frobenius 范数。

定义了矩阵范数就可以用来评估解线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 的适定性，即输入的 $\boldsymbol{b}$ 的误差导致的解 $\boldsymbol{x}$ 的误差大小。理论上，当 A 满秩时 $\boldsymbol{x}=A^{-1}\boldsymbol{b}$，设 $\boldsymbol{b}$ 有一个输入误差 $\Delta\boldsymbol{b}$，令

$$A\boldsymbol{x}=\boldsymbol{b},\quad A(\boldsymbol{x}+\Delta\boldsymbol{x}) = \boldsymbol{b}+\Delta\boldsymbol{b}, \tag{5.43}$$

则 $\boldsymbol{b}$ 输入的误差 $\Delta\boldsymbol{b}$ 造成的解的误差 $\Delta\boldsymbol{x}$ 满足 (见关治、陆金甫 (1998)[4] §5.4.2)

$$\frac{\|\Delta\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} \leq \kappa_p(A) \cdot \frac{\|\Delta\boldsymbol{b}\|_p}{\|\boldsymbol{b}\|_p}, \tag{5.44}$$

其中 $\kappa_p(A) = \|A\|_p \cdot \|A^{-1}\|_p$ 叫做 A 的**条件数**。条件数用来衡量以 A 为系数矩阵的线性方程组求解的稳定性，A 的条件数越大，求逆和解线性方程组越不稳定, 称系数矩阵条件数很大的线性方程组是**病态**的。

对于病态的线性方程组，某些变换可以解决一些明显的问题，比如，每一个方程都乘以一个调整因子使得各行元素大小相近，A 的每列都乘以一个调整因子使得各列元素大小相近（最后对应的未知数要作反向的调整），但是没有完全通用的方法解决病态问题。

求解线性方程组的误差还涉及到系数矩阵 A 的误差，更详细的讨论见关治、陆金甫 (1998)[4]§5.4.2。

5.3 线性方程组的特殊解法 *

对于特殊的系数矩阵，比如稀疏的、带状的、巨大的系数矩阵，需要利用其特殊结构以提高运算效率或减轻对存储空间要求。

5.3.1 带状矩阵

如果矩阵 A 的元素 a_{ij} 满足 $a_{ij} = 0$ 对 $i > j + p$ 和 $j > i + q$，则称 A 为**带状矩阵**，称 p 为下带宽，q 为上带宽。储存带状矩阵时可以排除零元素，仅保存其它元素，这样每行至多有 $p + q + 1$ 个元素，整个矩阵只保存 $n(p + q + 1)$ 个元素。带状矩阵如果保存在一个 $n \times (p + q + 1)$ 矩阵中，其 a_{ij} 元素在新的存储矩阵的 $(i, j - (i - p - 1))$ 位置。

线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 的系数矩阵如果是带状矩阵，用列主元的高斯消元法求解会在交换行时失去带状结构。如果 A 有 LU 分解 $A = LU$，易见下三角矩阵 L 是一个下带宽 p 的带状单位下三角矩阵，上三角矩阵 U 是上带宽 q 的带状上三角矩阵（见习题13）。

$p = q = 1$ 的带状矩阵叫做三对角矩阵，只需要把主对角线、下副对角线、上副对角线分

别保存在三个 n 维向量中。三对角矩阵如果有 LU 分解，必为如下形式:

$$M=\begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{pmatrix} \tag{5.45}$$

$$=\begin{pmatrix} 1 & & & & \\ l_2 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{n-1} & 1 & \\ & & & l_n & 1 \end{pmatrix}\begin{pmatrix} d_1 & c_1 & & & \\ & d_2 & c_2 & & \\ & & \ddots & \ddots & \\ & & & d_{n-1} & c_{n-1} \\ & & & & d_n \end{pmatrix} \tag{5.46}$$

所以只要求解 $l_2,\dots,l_n$ 和 $d_1,\dots,d_n$，输入 $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$ 三个向量后，求解得到 $l_2,\dots,l_n$ 可以放在 $a_2,\dots,a_n$ 的存储空间中，$d_1,\dots,d_n$ 放在 $b_1,\dots,b_n$ 的存储空间中，有如下的递推算法:

输入 $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$
for(i **in** $2:n$) {
  $a_i \leftarrow a_i/b_{i-1}$
  $b_i \leftarrow b_i - a_i c_{i-1}$
}
输出 $(a_2,\dots,a_n)$ 作为 $(l_2,\dots,l_n)$, $\boldsymbol{b}$ 作为 $\boldsymbol{d}$

三对角的线性方程组在 LU 分解后，用两次回代法解方程也可以利用 L 和 U 的带状结构简化算法。

4.2.2中求自然样条函数的未知参数 $M_1,M_2,\dots,M_n$ 的线性方程组就是三对角的。

带状的正定阵的 Cholesky 分解的结果也是带状的，可以简化计算。

例 5.3.1. 设 $\{\varepsilon_t,t\in\mathbb{Z}\}$ 是随机变量序列，其中 $\mathbb{Z}$ 表示所有整数组成的集合。$E\varepsilon_t\equiv 0$, $\mathrm{Var}(\varepsilon_t)\equiv\delta>0$, $\mathrm{cov}(\varepsilon_t,\varepsilon_s)=0$ 对 $s\neq t$，则称 $\{\varepsilon_t,t\in\mathbb{Z}\}$ 是白噪声列。若 $0<|b|<1$，

$$X_t=\varepsilon_t+b\varepsilon_{t-1},\ t\in\mathbb{Z}, \tag{5.47}$$

称 $\{X_t\}$ 为 MA(1) 序列。

设各 ε_t 服从 N(0,δ) 分布。若有 $\{X_t\}$ 的观测 $X_1,X_2,\dots,X_n$，则其联合分布完全依赖于 b,δ。令 $\boldsymbol{X}=(X_1,X_2,\dots,X_n)^T$，则 $E\boldsymbol{X}=\boldsymbol{0}$, $\mathrm{Var}(\boldsymbol{X})=\Sigma$, Σ 是一个三对角正定矩阵，

$\sigma_{ii}=\delta(1+b^2)$, $\sigma_{i+1,i}=\sigma_{i,i+1}=\delta b$。为了求 b,δ 的最大似然估计，注意到样本 $\boldsymbol{X}=\boldsymbol{x}$ 的对数似然函数为

$$\ln L(b,\delta)=-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\ln\det(\Sigma)-\frac{1}{2}\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}, \tag{5.48}$$

计算似然函数涉及给定参数 (b,δ) 后求相应的 Σ 的行列式以及计算二次型 $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}$，只要对 Σ 作 Cholesky 分解 $\Sigma=LL^T$, L 也是带状的，只有主对角线和下副对角线存在非零元素，L 满秩，这时 $\det(\Sigma)=[\det(L)]^2=(l_{11}l_{22}\dots l_{nn})^2$, $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}=(L^{-1}\boldsymbol{x})^T(L^{-1}\boldsymbol{x})$，只要用回代法解得 $\boldsymbol{y}=L^{-1}\boldsymbol{x}$ 则 $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}=\boldsymbol{y}^T\boldsymbol{y}$。

5.3.2 Toeplitz 矩阵

在时间序列分析中，若 $\{\xi_t,t\in\mathbb{Z}\}$ 是平稳时间序列，则 $\boldsymbol{\xi}=(\xi_1,\xi_2,\dots,\xi_n)^T$ 有协方差阵 $\Gamma_n=\big(\gamma(|i-j|)\big)_{n\times n}$, 其中 $\gamma(k)=\mathrm{cov}(\xi_1,\xi_{k+1})$ 叫做 $\{\xi_t\}$ 的协方差函数。这样形式的矩阵叫做 Toeplizt 矩阵，形式为

$$\Gamma_n=\begin{pmatrix}\gamma(0)&\gamma(1)&\cdots&\gamma(n-1)\\ \gamma(1)&\gamma(0)&\cdots&\gamma(n-2)\\ \vdots&\vdots&\ddots&\vdots\\ \gamma(n-1)&\gamma(n-2)&\cdots&\gamma(0)\end{pmatrix} \tag{5.49}$$

在很一般的条件下 Γ_n 正定。在本小节中假定 Γ_n 正定。如果要求解如下形式的方程，

$$\Gamma_n\boldsymbol{a}^{(n)}=\boldsymbol{\gamma}^{(n)}, \tag{5.50}$$

其中 $\boldsymbol{a}^{(n)}=(a_{n1},a_{n2},\dots,a_{nn})^T$, $\boldsymbol{\gamma}^{(n)}=(\gamma(1),\gamma(2),\dots,\gamma(n))^T$, 可以用一个 Levinson 递推算法求解，只需要 $O(n^2)$ 次运算 (见何书元 (2003) 2.4.3 小节)。$\boldsymbol{a}^{(n)}$ 叫做时间序列 $\{\xi_t\}$ 的 n 阶 Yule-Walker 系数，可用在 AR 建模或最优线性预报中。

在另外一些问题中可能需要对一般的 $\boldsymbol{y}^{(n)}$ 求解

$$\Gamma_n\boldsymbol{x}^{(n)}=\boldsymbol{y}^{(n)}, \tag{5.51}$$

其中 $\boldsymbol{x}^{(n)}=(x_{n1},x_{n2}\ \dots,x_{nn})^T$, $\boldsymbol{y}^{(n)}=(y_1,y_2,\dots,y_n)^T$, 比如计算 $\boldsymbol{\xi}$ 的似然函数时。(5.50)是(5.51)的特例。下面利用 Toeplitz 矩阵的特殊结构构造(5.51)的高效算法。

首先就考虑 Y-W 方程(5.50)的求解。令 P_n 表示把 $(1,2,\dots,n)$ 排列为 $(n,n-1,\dots,1)$ 的置换阵,则 $P_nP_n=I_n$, $P_n\Gamma_n^{-1}=\Gamma_n^{-1}P_n$, $P_n\Gamma_n^{-1}P_n=\Gamma_n^{-1}$。记 $\boldsymbol{a}_*^{(n+1)}=(a_{n+1,1},a_{n+1,2},\dots,a_{n+1,n})^T$,

则 $n+1$ 阶的 Y-W 方程可以写成如下分块形式

$$\begin{pmatrix} \Gamma_n & P_n\boldsymbol{\gamma}^{(n)} \\ \boldsymbol{\gamma}^{(n)T}P_n & \gamma(0) \end{pmatrix} \begin{pmatrix} \boldsymbol{a}_*^{(n+1)} \\ a_{n+1,n+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma}^{(n)} \\ \gamma(n+1) \end{pmatrix} \tag{5.52}$$

利用矩阵消元的想法把上面的第 $n+1$ 个方程中的 $\boldsymbol{\gamma}^{(n)T}P_n$ 变成零，只要用 $-\boldsymbol{\gamma}^{(n)T}P_n\Gamma_n^{-1}$ 左乘前 n 个方程然后加到第 $n+1$ 个方程即可，这时有

$$\Gamma_n a_*^{(n+1)} + a_{n+1,n+1}P_n\gamma^{(n)} = \boldsymbol{\gamma}^{(n)}, \tag{5.53}$$

$$\left(\gamma(0) - \boldsymbol{\gamma}^{(n)T}P_n\Gamma_n^{-1}P_n\boldsymbol{\gamma}^{(n)}\right) a_{n+1,n+1} = \gamma(n+1) - \boldsymbol{\gamma}^{(n)T}P_n\Gamma_n^{-1}\boldsymbol{\gamma}^{(n)}, \tag{5.54}$$

于是

$$a_{n+1,n+1} = \frac{\gamma(n+1) - \boldsymbol{\gamma}^{(n)T}P_n\boldsymbol{a}^{(n)}}{\gamma(0) - \boldsymbol{\gamma}^{(n)T}\boldsymbol{a}^{(n)}} = \frac{\gamma(n+1) - a_{n1}\gamma(n) - \cdots - a_{nn}\gamma(1)}{\gamma(0) - a_{n1}\gamma(1) - \cdots - a_{nn}\gamma(n)}, \tag{5.55}$$

$$\boldsymbol{a}_*^{(n+1)} = \boldsymbol{a}^{(n)} - a_{n+1,n+1}P_n\boldsymbol{a}^{(n)}. \tag{5.56}$$

为求解(5.51)，类似对 $n+1$ 阶的方程进行分块消元，分块形式为

$$\begin{pmatrix} \Gamma_n & P_n\boldsymbol{\gamma}^{(n)} \\ \boldsymbol{\gamma}^{(n)T}P_n & \gamma(0) \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_*^{(n+1)} \\ x_{n+1,n+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}^{(n)} \\ y_{n+1} \end{pmatrix} \tag{5.57}$$

其中 $\boldsymbol{x}_*^{(n+1)}$ 是 $\boldsymbol{x}^{(n+1)}$ 的前 n 个元素组成的列向量。消去第 $n+1$ 个方程中的 $\boldsymbol{\gamma}^{(n)T}P_n$，得

$$x_{n+1,n+1} = \frac{y_{n+1} - \boldsymbol{\gamma}^{(n)T}P_n\boldsymbol{x}^{(n)}}{\gamma(0) - \boldsymbol{\gamma}^{(n)T}\boldsymbol{a}^{(n)}} = \frac{y_{n+1} - x_{n1}\gamma(n) - \cdots - x_{nn}\gamma(1)}{\gamma(0) - a_{n1}\gamma(1) - \cdots - a_{nn}\gamma(n)}, \tag{5.58}$$

$$\boldsymbol{x}_*^{(n+1)} = \boldsymbol{x}^{(n)} - x_{n+1,n+1}P_n\boldsymbol{a}^{(n)}. \tag{5.59}$$

为求解 $\boldsymbol{x}^{(n)}$ 需要同时计算 $\boldsymbol{a}^{(n)}$，总共只需要 $O(n^2)$ 次运算。

5.3.3 稀疏系数矩阵方程组求解

系数矩阵为带状矩阵时解方程组计算量可以从 $O(n^3)$ 降低到 $O(n)$。更一般的稀疏矩阵的非零元素分布不一定有固定的规律，而且在求解消元过程中可能变得不再稀疏。现代统计模型中经常有数千自变量的情形，涉及的矩阵经常是稀疏矩阵，减少不必要的存储并加速运算可以使这样的问题求解变得可行或者更加高效。在经典的统计问题中，有许多个因素的试验的设计阵是稀疏矩阵。

稀疏矩阵可以只保存非零元素与其所在的行列位置。在设计存储方案时，要考虑到如何使用此矩阵，比如，仅用来计算 $A\boldsymbol{x}$ 这样的矩阵和向量的乘法，要求解线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$，矩阵是否会添加、删除元素或修改元素值，访问矩阵时需要按行访问还是按列访问，等等。

下面假设对稀疏矩阵 A 主要需要计算 $\boldsymbol{y} \leftarrow A\boldsymbol{x}$ 这样的矩阵和向量的乘法。如果 $A, \boldsymbol{x}, \boldsymbol{y}$ 都可以保存到内存中，不需要修改 A 的内容，则可以把 A 的非零元素存入一个长度为 m 的数组 $\boldsymbol{a}$(m 是 A 的非零元素个数), 把非零元素的行号存入长度为 m 的数组 $\boldsymbol{r}$, 列号存入数组 $\boldsymbol{c}$，即 $a_{ij} \neq 0$ 保存在 $\boldsymbol{a}$ 的第 k 元素中，则 $r_k = i$, $c_k = j$。乘法 $\boldsymbol{y} \leftarrow A\boldsymbol{x}$ 的伪代码为:

输入 $\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{c}, \boldsymbol{x}$
$\boldsymbol{y} \leftarrow \boldsymbol{0}$
for(k **in** $1 : m$) { # 计算与第 k 个 A 非零元素有关的乘法
  $i \leftarrow r_k$, $j \leftarrow c_k$
  $y_i \leftarrow y_i + a_k x_j$
} # end for k
输出 $\boldsymbol{y}$

以上算法稍作修改就可以适用于 A 的所有非零元素无法同时调入内存，需要分批调入的情形。

如果 A 可能被修改，就要设计存储方法使得能迅速定位元素。比如，可以把 A 的非零元素按行存储为链表，然后保存每行的非零元素个数，并把每行的非零元素所在的列号保存为链表。这样可以在一行中迅速找到某个元素的值, 也可以删除非零元素、添加非零元素、修改元素值。

5.3.4 用迭代法求解线性方程组

用消元法求解 $A\boldsymbol{x} = \boldsymbol{b}$ 需要 $O(n^3)$ 次运算, 可以在有限步结束。迭代算法每次给出解 $\boldsymbol{x}$ 的一个近似，下次迭代对此近似进行改进，每次迭代仅需要 $O(n^2)$ 次运算，当达到需要的精度时停止迭代。如果迭代很快收敛，这种方法可以比消元法更快求解。如果 A 是稀疏矩阵，消元法会在消元过程中使矩阵变得稠密，而迭代法则可以保持使用稀疏矩阵，从而减少计算量。

Jacobi 方法是求解线性方程组的一种迭代算法。对线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$, 首先把 A 分解为 $A = L + D + U$, 其中 L 仅有 A 的严格下三角部分，D 为 A 的主对角部分，U 仅有 A 的严格上三角部分。设迭代的第 k 步已经得到了近似解 $\boldsymbol{x}^{(k)}$, 则在第 $k+1$ 步时对 $i = 1, 2, \ldots, n$ 计算

$$x_i^{(k+1)} = \frac{b_i - \sum_{j \neq i} a_{ij} x_j^{(k)}}{a_{ii}} \tag{5.60}$$

写成向量形式为

$$\boldsymbol{x}^{(k+1)} = D^{-1} \left[\boldsymbol{b} - (L + U)\boldsymbol{x}^{(k)} \right]. \tag{5.61}$$

在迭代收敛时，(5.61)变成

$$D\boldsymbol{x} = \boldsymbol{b} - L\boldsymbol{x} - U\boldsymbol{x} \Longleftrightarrow (L + D + U)\boldsymbol{x} = \boldsymbol{b} \tag{5.62}$$

即 $A\boldsymbol{x} = \boldsymbol{b}$。

显然，Jacobi 方法要求 A 的主对角线元素都不等于零。为了使得迭代收敛，即 $\lim_{k\to\infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| = 0$, 注意到第 $k+1$ 步的误差为

$$\left(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\right) = -D^{-1}(L + U)\left(\boldsymbol{x}^{(k)} - \boldsymbol{x}\right) = C^{k+1}(\boldsymbol{x}^{(0)} - \boldsymbol{x}),$$

其中 $C = D^{-1}(L + U)$，所以迭代收敛的充分条件为所有 $a_{ii} \neq 0$ 且 $\|C\| < 1$，这时

$$\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}\| \leq \|C\|^{k+1}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}\| \to 0 \quad (k \to \infty).$$

其中 $\|C\|$ 是 C 的任意一种矩阵范数。可以证明，在所有 $a_{ii} \neq 0$ 时，Jacobi 算法收敛的充分必要条件为 $\|C\|_2 < 1$ (参见关治、陆金甫 (1998)[4]§6.1)。

在某些问题中以上收敛性条件可以得到验证。为了判断迭代是否已经收敛，一般预先指定一个精度 ϵ, 当 $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| < \epsilon$ 时停止迭代。由于舍入误差的影响，ϵ 只要取为问题需要的精度即可，取过小的 ϵ 有可能会造成算法无法停止。

以上的 Jacobi 方法在第 $k+1$ 步利用 $\boldsymbol{x}^{(k)}$ 得到整个 $\boldsymbol{x}^{(k+1)}$。另外一种迭代方法叫做 Gauss-Seidel 迭代，每次仅更新一个分量，而且更新后的分量值马上在下一步中就可以利用。迭代公式为

$$x_i^{(k+1)} = \frac{b_i - \sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)}}{a_{ii}}, \tag{5.63}$$

写成向量形式为

$$(L + D)\boldsymbol{x}^{(k+1)} = \boldsymbol{b} - U\boldsymbol{x}^{(k)} \Longleftrightarrow \boldsymbol{x}^{(k+1)} = (L + D)^{-1}(\boldsymbol{b} - U\boldsymbol{x}^{(k)}),$$

这种方法收敛的充分必要条件为 $L + D$ 可逆且 $\|(L + D)^{-1}U\|_2 < 1$。特别地，当 A 为对称正定阵时 Gauss-Seidel 迭代必定收敛 (参见关治、陆金甫 (1998)[4]§6.2)。

对于 Gauss-Seidel 方法，一种改进的方法是 $x_i^{(k+1)}$ 取为(5.63)与 $x_i^{(k)}$ 的加权平均，这种方法称为超松弛 (SOR) 迭代法，某些特殊的系数矩阵可以找到最优的加权因子使得收敛速度达到最优。

当 A 为对称正定阵时，$A\boldsymbol{x} = \boldsymbol{b}$ 的解等价于 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$ 的最小值点，可以用函数最优化方法如共轭梯度法求解。

5.4 QR 分解

考虑如下线性模型的估计问题:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.64}$$

其中 X 是 $n \times p$ 已知矩阵 $(n > p)$，$\boldsymbol{y}$ 是 n 维因变量观测值向量，$\boldsymbol{\beta}$ 是 p 为未知模型系数向量，$\boldsymbol{\varepsilon}$ 是 n 维未知的模型随机误差向量。用最小二乘法估计 $\boldsymbol{\beta}$，即求 $\boldsymbol{\beta}$ 使得

$$g(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \tag{5.65}$$

最小 (本节中的向量范数 $\|\boldsymbol{x}\|$ 都表示欧式空间中的长度 $\sqrt{\boldsymbol{x}^T\boldsymbol{x}}$)，归结为求解如下正规方程

$$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}. \tag{5.66}$$

此方程一定有解，当 X 列满秩时有唯一解 $\hat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\boldsymbol{y}$。

考虑正规方程(5.66)的数值计算问题。我们需要解出 $\boldsymbol{\beta}$(记作 $\hat{\boldsymbol{\beta}}$), 并计算残差平方和

$$\text{SSE} = \|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2. \tag{5.67}$$

在 X 列满秩时，正规方程(5.66)的系数矩阵 $X^T X$ 是正定阵，可以用 Cholesky 分解方法高效地计算出 $\hat{\boldsymbol{\beta}}$ 和 SSE（参见习题11），并且所需的存储空间也只有 $O(p^2)$ 阶。但是，当 $X^T X$ 条件数 $\kappa_2(X^T X)$ 很大时，直接求解正规方程(5.66)会引入很大误差。有结果表明（参见 Stewart(1973)[36] 定理 5.2.4），$\hat{\boldsymbol{\beta}}$ 的相对误差有如下控制式:

$$\frac{\|\Delta\hat{\boldsymbol{\beta}}\|}{\|\hat{\boldsymbol{\beta}}\|} \leq \sqrt{\kappa_2(X^T X)}\frac{\|\Delta\hat{\boldsymbol{y}}\|}{\|\hat{\boldsymbol{y}}\|}, \tag{5.68}$$

其中 $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$, $\Delta\hat{\boldsymbol{y}}$ 是观测值 $\boldsymbol{y}$ 中的误差引起的 $\hat{\boldsymbol{y}}$ 的误差，$\Delta\hat{\boldsymbol{\beta}}$ 是观测值 $\boldsymbol{y}$ 中的误差引起的 $\hat{\boldsymbol{\beta}}$ 的误差。所以，应该寻找计算更稳定的算法来求解最小二乘问题(5.65)。另外，在样本量 n 很大时，计算 $X^T X$ 时的累积误差可能导致其实际不正定，使得 Cholesky 分解算法失败。新的算法最好能够在 X 非列满秩时也可以求出适当的最小二乘解。

注意，在统计数据分析中，数据中的随机误差和舍入误差往往超过计算中的误差，但不稳定的算法会放大这些数据中的误差。一般只要控制计算中造成的误差比随机误差小一个数量级就可以满足要求。

5.4.1 Gram-Schmidt 正交化方法

定义 (QR 分解) 对 $n\times p$ 矩阵 X，若有 $n\times p$ 矩阵 Q 满足 $Q^TQ=I_p$，以及 p 阶上三角矩阵 R 使得

$$X=QR, \tag{5.69}$$

则称(5.69)为矩阵 X 的 QR 分解，或正交—三角分解。

假设最小二乘问题(5.65)中的 X 有 QR 分解 $X=QR$ 且 R 满秩。设 $Q^*=(Q\,|\,Q_2)$ 是 n 阶正交阵，则

$$\|\boldsymbol{y}-X\boldsymbol{\beta}\|^2=\|(Q^*)^T(\boldsymbol{y}-QR\boldsymbol{\beta})\|^2=\|Q^T\boldsymbol{y}-R\boldsymbol{\beta}\|^2+\|Q_2^T\boldsymbol{y}\|^2, \tag{5.70}$$

$\|\boldsymbol{y}-X\boldsymbol{\beta}\|^2$ 与 $\|Q^T\boldsymbol{y}-R\boldsymbol{\beta}\|^2$ 有相同的最小值点，用回代法求解

$$R\boldsymbol{\beta}=Q^T\boldsymbol{y} \tag{5.71}$$

就可以得到最小二乘估计 $\hat{\boldsymbol{\beta}}$, 而(5.70)式右边的 $\|Q_2^T\boldsymbol{y}\|^2$ 就是残差平方和。由于 $X^TX=R^TR$，R^T 是下三角形矩阵，所以从 QR 分解也可以得到 Cholesky 分解 (可能符号有差别)。

(5.69)实际是把 X 的各列进行了正交化。把 X 的第一列标准化为 Q 的第一列，把 X 的第一、二列正交化和标准化得到 Q 的第二列，以此类推，这正是线性代数中的 Gram-Schmidt 正交化过程, 算法用伪代码表示如下:

令 $r_{11}\leftarrow\|X_{\cdot1}\|$, $Q_{\cdot1}\leftarrow r_{11}^{-1}X_{\cdot1}$.
for(j **in** $2:p$) { # 求解 $Q_{\cdot j}$ 和 R 的第 j 列
  $Q_{\cdot j}\leftarrow X_{\cdot j}$
  for(k **in** $1:(j-1)$){
    $r_{kj}\leftarrow X_{\cdot j}^TQ_{\cdot k}$
    $Q_{\cdot j}\leftarrow Q_{\cdot j}-r_{kj}Q_{\cdot k}$
  }
  $r_{jj}\leftarrow\|\boldsymbol{Q}_{\cdot j}\|$, $Q_{\cdot j}\leftarrow r_{jj}^{-1}Q_{\cdot j}$
} # end for j
输出 Q 和 R

以上的 Gram-Schmidt 算法 (记作 RGS) 虽然得到了 X^TX 的 QR 分解，但是该算法得到的 Q 矩阵由于计算误差影响可能会正交性不够好，使得用这样的到 Q 矩阵以及(5.71)求解 $\hat{\boldsymbol{\beta}}$ 的条件数与直接求解正规方程(5.66)的条件数相同。对 Gram-Schmidt 算法略作修改就可以得到更高精度的分解结果。

修正的 Gram-Schmidt 算法 (记作 MGS) 仅仅调整了计算的次序，并把 Q 的分解结果保存在了 X 的存储空间中。在 MGS 的第 j 步计算中求 Q 的第 j 列和 R 的第 j 行，把 Q 的第 j 列保存在 X 的第 j 列中，并把 X 的后续列减去其在 $Q_{\cdot j}$ 上的投影。算法用伪代码表示如下:

for(j **in** $1:p$) { # 求解 $Q_{\cdot j}$ 和 R 的第 j 行
  $r_{jj} \leftarrow \|X_{\cdot j}\|$, $X_{\cdot j} \leftarrow r_{jj}^{-1} X_{\cdot j}$
  for(k **in** $(j+1):n$){
    $r_{jk} \leftarrow X_{\cdot j}^T X_{\cdot k}$
    $X_{\cdot k} \leftarrow X_{\cdot k} - r_{jk} X_{\cdot j}$
  }
} # end for j
输出 $Q = X$ 和 R

在以上的 RGS 和 MGS 算法过程中，r_{jj} 是 X 的第 j 列对 X 的第 $1, 2, \ldots, j-1$ 列作线性回归 (无截距项) 得到的残差平方和的平方根, 所以某一步中如果 $r_{jj} = 0$ 说明 X 的第 j 列与前面的列共线。在 MGS 算法过程中，第 j 步以后，X 的第 $j+1, j+2, \ldots, p$ 列中保存的是原始的 X 的那些列关于 Q 的前 j 列回归 (无截距项) 后的残差, 这时类似于解线性方程组时的主元方法, 可以优先选后续列中向量范数最大的列对应的自变量进入模型。

在 X 列满秩时，RGS 和 MGS 算法得到的上三角阵 R 的主对角线元素都为正值，所以 $X^T X = R^T R$ 是 $X^T X$ 的 Cholesky 分解。

5.4.2 Householder 变换 *

Householder 变换方法逐次对 X 进行正交变换把 X 变成上三角形，这样就得到了 X 的 QR 分解。

设 $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)^T$ 是一个 m 维向量，找一个正交变换把 $\boldsymbol{x}$ 最后 $m-1$ 个元素都变成零。取 $U = I_m - d\boldsymbol{u}\boldsymbol{u}^T$, 其中 $d = 2/\boldsymbol{u}^T\boldsymbol{u}$，易见对任意的 m 维非零向量 $\boldsymbol{u}$，U 都是对称的正交方阵 ($U^T U = UU^T = I_m$), 且对任意常数 b，$b\boldsymbol{u}$ 和 $\boldsymbol{u}$ 对应相同的 U 矩阵。给定向量 $\boldsymbol{x}$ 后，来求 $\boldsymbol{u}$ 使得 $U\boldsymbol{x}$ 仅有第一个元素非零，即 $U\boldsymbol{x} = s\boldsymbol{e}_1$($\boldsymbol{e}_1$ 是仅有第一个元素为 1，其它元素都等于 0 的 m 维向量)，由于正交变换是使得向量长度不变的，有 $\|\boldsymbol{x}\|^2 = \|U\boldsymbol{x}\|^2 = s^2\|\boldsymbol{e}_1\|^2 = s^2$, 即 $s^2 = \|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x}$。由 $U\boldsymbol{x} = s\boldsymbol{e}_1$ 得 $(d\boldsymbol{u}^T\boldsymbol{x})\boldsymbol{u} = \boldsymbol{x} - s\boldsymbol{e}_1$, 即 $\boldsymbol{u}$ 是 $\boldsymbol{x} - s\boldsymbol{e}_1$ 的数乘结果，因为 $\boldsymbol{u}$ 和 $\boldsymbol{u}$ 的非零数乘结果对应相同的矩阵 U, 所以不妨取 $\boldsymbol{u} = \boldsymbol{x} - s\boldsymbol{e}_1$，其中 $s = \|\boldsymbol{x}\|$。容易验证，这样选取的 u 使得 $U\boldsymbol{x} = \|\boldsymbol{x}\|\boldsymbol{e}_1$。对向量 $\boldsymbol{x}$ 的这种正交变换叫做 Householder **变换**。

可以看出，对任何的 m 维向量 $\boldsymbol{x}$ 都可以用 Householder 变换把最后 $m-1$ 个元素都变成零。对一个 $n\times p$ 矩阵 $X(n>p)$, 可以左乘 $X_{\cdot 1}$ 的 Householder 变换阵，把 $X_{\cdot 1}$ 的最后 $n-1$ 个元素变成零；然后对变换后的 X, 左乘一个变换矩阵，使得 X 的第 1 行不变，而使得 $X_{\cdot 2}$ 的最后 $n-2$ 个元素变成零，即仅对 X 的第 2 到 n 行作 $n-1$ 维的 Householder 变换，而且第一列中的最后 $n-1$ 个零不变。以此类推，在第 j 次变换时仅对 X 的第 j 到 n 行作 $n-j+1$ 维的 Householder 变换。于是，变换 p 次后，X 变成了一个上三角矩阵 R ($n\times p$ 的上三角矩阵是指当 $i>j$ 时总有 $r_{ij}=0$)。

令 $X^{(0)}=X$,

$$U_j=\begin{pmatrix} I_{j-1} & \mathbf{0} \\ \mathbf{0} & U^{(j)} \end{pmatrix}, \quad X^{(j)}=U_jX^{(j-1)}, \tag{5.72}$$

其中 U_j 是 $X_{\cdot j}^{(j)}$ 的最后 $n-j+1$ 个元素的 Householder 变换矩阵。用 $\boldsymbol{x}^{(j)}$ 表示 $X_{\cdot j}^{(j)}$ 的最后 $n-j+1$ 个元素，令 $s_j=\|x^{(j)}\|$, $\boldsymbol{u}^{(j)}=\boldsymbol{x}^{(j)}-s_j\boldsymbol{e}_1^{(n-j+1)}$, $U^{(j)}=I_{n-j+1}-\frac{2}{\|\boldsymbol{u}^{(j)}\|^2}\boldsymbol{u}^{(j)}(\boldsymbol{u}^{(j)})^T$, 其中 $e_1^{(n-j+1)}$ 表示仅有第一个元素等于 1，所有其它元素等于 0 的 $n+1-j$ 维向量。在 p 步变换后得到

$$R^*=U_pU_{p-1}\ldots U_1X \tag{5.73}$$

是一个 $n\times p$ 的上三角阵，设 R^* 的前 p 行组成的矩阵为 R, 则 R 是 p 阶上三角方阵，而 R^* 的后 $n-p$ 行均为零。令 $U^*=U_pU_{p-1}\cdot U_1$, 则 U^* 是一个 n 阶对称正交阵, 设 U^* 的前 p 列组成的 $n\times p$ 矩阵为 U，则

$$X=U^*R^*=(U\ \ *)\begin{pmatrix} R \\ \mathbf{0} \end{pmatrix}=UR, \tag{5.74}$$

于是用 p 次 Householder 变换得到了 X 的 QR 分解。用 Householder 变换作 QR 分解来求解最小二乘问题是计算精度较高的方法。

编程实现时，在 X 的存储空间中保存每次得到的 $X^{(j)}$，因为第 j 列的最后 $n-j$ 个元素为零，第 j 个元素为 s_j, 所以可以单独保存 $s_1,s_2,\ldots,s_p$, 并把 $\boldsymbol{u}^{(j)}$ 保存在 X 的第 j 列的最后 $n-j+1$ 个元素的存储空间中。

当 X 列满秩时，R 的主对角线元素都为正值，$X^TX=R^TR$ 是 X^TX 的 Cholesky 分解。

如果 X 存在共线问题，比如，$X_{\cdot j}$ 可以用 X 的前 $j-1$ 列线性表示，则 Householder 变换过程进行了 $j-1$ 次以后，因为 X 的前 $j-1$ 列的最后 $n-j+1$ 行的元素都变成了零，所以 X 的第 j 列的最后 $n-j+1$ 个元素也都变成了零，这时 $s_j=\|\boldsymbol{x}^{(j)}\|=0$。这样在回归计算中可以判断共线是否发生，也可以用类似选主元的方法，在第 j 步看后面的 $n-j+1$ 列的

后 $n-j+1$ 行组成的子矩阵中哪一列的向量范数最大，就把那一列对应的自变量与第 j 个自变量交换位置。

5.4.3 Givens 变换 *

Givens 变换也是一种正交变换，是二维向量的旋转变换的推广，可以把向量的指定元素变成零。

对二维非零向量 $\boldsymbol{x}=(x_1,x_2)^T$, 把 $\boldsymbol{x}$ 看成直角坐标平面上的向量，作旋转变换把新的 x 轴旋转到 $\boldsymbol{x}$ 的方向上，设旋转角度为 θ, 则

$$\cos\theta=\frac{x_1}{\sqrt{x_1^2+x_2^2}},\quad \sin\theta=\frac{x_2}{\sqrt{x_1^2+x_2^2}}, \tag{5.75}$$

旋转变换矩阵和变换结果为

$$R=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix}=(x_1^2+x_2^2)^{-\frac{1}{2}}\begin{pmatrix}x_1 & x_2\\ -x_2 & x_1\end{pmatrix},\quad R\boldsymbol{x}=\begin{pmatrix}\sqrt{x_1^2+x_2^2}\\ 0\end{pmatrix}. \tag{5.76}$$

对 n 维向量 $\boldsymbol{x}$, 为了把 x_k 变成零，可以对 $(x_i,x_k)(i<k)$ 作旋转使得 x_k 变成零，保持其它元素不变。这相当于左乘了一个正交变换矩阵 U_{ik}, U_{ik} 与单位阵 I_n 仅在 4 个元素上有差别: 其 $(i,i),(i,k),(k,i),(k,k)$ 元素组成的子矩阵恰好为 (x_i,x_k) 对应的旋转阵

$$R_{ik}=(x_i^2+x_k^2)^{-\frac{1}{2}}\begin{pmatrix}x_i & x_k\\ -x_k & x_i\end{pmatrix}, \tag{5.77}$$

则 $U_{ik}\boldsymbol{x}$ 除去第 i,k 号元素以外不变，第 i 号元素变成 $\sqrt{x_i^2+x_k^2}$, 第 k 号元素变成 0。

仿照用 Householder 变换进行 QR 分解的方法，利用 Givens 变换也可以对 $n\times p$ 矩阵 $X(n>p)$ 作 QR 分解。首先，用 $n-1$ 次 Givens 变换可以把 X 的第 1 列的最后 $n-1$ 个元素变成零，每次用该列主对角线元素与其下面的元素构成 Givens 变换。然后，用 $n-2$ 次 Givens 变换可以把 X 的第 2 列的最后 $n-2$ 个元素变成零，每次用该列主对角线元素与其下面的一个元素构成 Givens 变换。设已经把 X 的前 $j-1$ 列变成了上三角形 (前 $j-1$ 列主对角线下方的元素都变成了零), 在第 j 步，用 $n-j$ 次 Givens 变换把主对角线下方的元素都变成零，每次用该列主对角先元素与其下面的一个元素构成 Givens 变换，注意此变换仅影响到第 j 行和第 k 行 $(k>j)$, 所以不会影响到所有列的前 $j-1$ 行元素，也就不会影响到前 $j-1$ 列的上三角部分，前 $j-1$ 列已经变成零的那些元素也保持为零，第 j 列中已经变成零的元素也保持为零。注意，第 j 步关于第 j 和第 k 元素作 Givens 变换时，仅需对当前 X 的第 j,k 两行和第 $j,j+1,\ldots,p$ 列组成的两行的子矩阵作二维的 Givens 变换即可, 并且对第 j

列的两个元素的变换结果分别为 $\sqrt{x_j^2+x_k^2}$ 和 0，不用重复计算。如此进行 p 步后就把 X 变成了上三角形，共需要进行 $(n-1)+(n-2)+\cdots+(n-p)=np-\frac{1}{2}p(p+1)$ 次 Givens 变换。

用 Givens 变换作 QR 分解，还可以按照每次消除主对角下方的一行的方式: 首先消除主对角线下方第 2 行 (只有 $(2,1)$ 号元素), 然后消除主对角线下方第 3 行 (有 $(3,1,),(3,2)$ 两个元素), 如此重复直到消除了主对角线下方第 n 行元素 (该行所有元素)。容易看出，按这种次序变换，在消去第 $i+1$ 行下三角部分的元素时，不会影响前面的第 $1,2,\ldots,i$ 行中下三角部分已经变成零的元素，但是会改变第 $1,2,\ldots,i$ 行的上三角部分 (包含对角线) 的元素。在回归计算中采用这种变换次序可以适应不断获得新的观测需要更新回归结果的情况（参见 Monahan(2001)[31] §5.8）。

借助于 QR 分解可以很容易地计算帽子矩阵 $X(X^TX)^{-1}X^T$ 的主对角线元素 h_i 的值，并由此计算多种回归诊断统计量，比如外学生化残差。回归的一般线性约束的假设检验也可以借助于正交分解的方法计算。参见 Monahan(2001)[31] §5.9 和 §5.10。**消去变换** (sweep) 是解线性方程组时完全消元法（把系数矩阵通过初等变换化为单位阵）的变种，在原来系数矩阵的存储空间中保存得到的逆矩阵，为回归变量选择的计算提供了很多便利，参见高惠璇 (1995)[2] §5.6 和 Monahan(2001)[31] §5.12。

在 R 软件中，用 `qr()` 函数可以计算矩阵的 QR 分解。

5.5 特征值、奇异值

5.5.1 定义

在多元统计和时间序列分析中会用到特征值和奇异值, 比如，主成分分析、典型相关分析、多元自回归模型等。

先简单回顾线性代数中特征值的定义和性质。设 A 为 n 阶方阵，若有非零向量 $\boldsymbol{\alpha}$ 和复数 λ 使得

$$A\boldsymbol{\alpha}=\lambda\boldsymbol{\alpha}, \tag{5.78}$$

则称 λ 是矩阵 A 的一个**特征值**，$\boldsymbol{\alpha}$ 是特征值 λ 对应的**特征向量**。当 A 为 n 阶实对称阵时，A 恰有 n 个实特征值，记作 $\lambda_1,\lambda_2,\ldots,\lambda_n$, 相应的特征向量也都是实向量, 且存在 n 阶正交阵 U 使得

$$A=U\Lambda U^T=\sum_{j=1}^{n}\lambda_j\boldsymbol{u}_{\cdot j}\boldsymbol{u}_{\cdot j}^T, \tag{5.79}$$

其中 $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ 是对角线元素为 A 的特征值的对角阵, U 的第 j 列 $\boldsymbol{u}_{\cdot j}$ 是 λ_j 对应的特征向量。当 A 为非负定阵时，所有特征值非负；当 A 为正定阵时，所有特征值都是正数。

在统计问题如典型相关分析中还会遇到广义的特征值问题: 设 A, B 为 n 阶方阵，若有复数 λ 和非零向量 $\boldsymbol{\alpha}$ 使得

$$A\boldsymbol{\alpha} = \lambda B\boldsymbol{\alpha}, \tag{5.80}$$

则称 λ 和 $\boldsymbol{\alpha}$ 分别为矩阵 A 相对于矩阵 B 的**广义特征值**和**广义特征向量**。实际问题中，B 通常是正定阵，A 是实对称阵，这时(5.80)等价于 $B^{-1}A\boldsymbol{\alpha} = \lambda\boldsymbol{\alpha}$，可以化为普通特征值问题，并可利用 Cholesky 分解进行计算。设 B 有 Cholesky 分解 $B = LL^T$, 则由 $A\boldsymbol{\alpha} = \lambda LL^T\boldsymbol{\alpha}$ 得 $L^{-1}A(L^T)^{-1}(L^T\boldsymbol{\alpha}) = \lambda(L^T\boldsymbol{\alpha})$, 求解普通特征值问题 $\left(L^{-1}A(L^T)^{-1}\right)\boldsymbol{\beta} = \lambda\boldsymbol{\beta}$ 得 λ 和 $\boldsymbol{\beta}$ 再求解 $L^T\boldsymbol{\alpha} = \boldsymbol{\beta}$ 即可得广义特征值和广义特征向量。

对 n 阶非奇异矩阵 A, 必存在 n 阶正交阵 U 和 V，使得

$$A = V\,\mathrm{diag}(d_1, d_2, \ldots, d_n)U^T, \tag{5.81}$$

其中 $d_1, d_2, \ldots, d_n$ 是正定阵 A^TA 的 n 个特征值的算数平方根, 称(5.81)为矩阵 A 的**奇异值分解**, $d_1, d_2, \ldots, d_n$ 称为 A 的**奇异值**。

若 A 是一般的 $n \times m$ 非零矩阵，A 的秩为 $\mathrm{rank}(A) = r \le \min(n, m)$, A^TA 的非零特征值为 $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r > 0$, 令 $d_i = \sqrt{\lambda_i}$, $i = 1, 2, \ldots, r$, 则称 d_i 为 A 的**奇异值**, 且一定有 m 阶正交阵 U 和 n 阶正交阵 V 使得

$$A = VDU^T, \tag{5.82}$$

其中

$$D = \begin{pmatrix} \mathrm{diag}(d_1, d_2, \ldots, d_r) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}_{n\times m}, \tag{5.83}$$

称(5.82)为 A 的**奇异值分解** (singular value decomposition, SVD)。详见高惠璇 (1995)[2]§5.4 和 Monahan(2001)[31] §6.6。

5.5.2 对称阵特征值分解的 Jacobi 算法 *

矩阵 A 的特征值 λ 是 A 的特征多项式 $A - \lambda I$ 的根，但直接求多项式的根并不容易，特征值和特征向量的计算一般都通过迭代算法实现。

§5.4.3引入的 Givens 变换是一个旋转变换，可以仅改变向量中指定的两个元素并使得第二个指定元素变成零。类似这样仅改变向量中第 i,j 两个元素的旋转变换矩阵可以写成

$$G_{ij}(\theta)=\begin{pmatrix}
1 & & & & & & & & & \\
& \ddots & & & & & & & & \\
& & 1 & & & & & & & \\
& & & \cos\theta & & & \sin\theta & & & \\
& & & & 1 & & & & & \\
& & & & & \ddots & & & & \\
& & & & & & 1 & & & \\
& & & -\sin\theta & & & \cos\theta & & & \\
& & & & & & & 1 & & \\
& & & & & & & & \ddots & \\
& & & & & & & & & 1
\end{pmatrix}. \tag{5.84}$$

若 A 为对称阵，适当选取角度 θ 对 A 作如下变换

$$A^*=G_{ij}(\theta)A\left(G_{ij}(\theta)\right)^T \tag{5.85}$$

可以使得 $a^*_{ij}=a^*_{ji}=0$，这样的变换叫做 Jacobi 变换。对 A 反复地作 Jocobi 变换可以使得非对角线元素趋于零。

考虑 Jacobi 变换(5.85)中角度 θ 的确定。显然，A^* 和 A 的不同仅体现在第 i,j 行和第 i,j 列，其它元素保持不变; $G_{ij}(\theta)A$ 与 A 仅在第 i,j 行有差别，$A\left(G_{ij}(\theta)\right)^T$ 与 A 仅在第 i,j 列有差别，A^* 与 $G_{ij}(\theta)A$ 仅在第 i,j 列有差别。简单推导可得

$$a^*_{ij}=\frac{1}{2}(a_{jj}-a_{ii})\sin 2\theta+a_{ij}\cos 2\theta, \tag{5.86}$$

只要取 $\theta\in(-\frac{\pi}{4},\frac{\pi}{4})$ 使得

$$\tau\triangleq\cot 2\theta=\frac{a_{ii}-a_{jj}}{2a_{ij}} \tag{5.87}$$

即可使 $a^*_{ij}=a^*_{ji}=0$。

注意到 $G_{ij}(\theta)$ 仅依赖于 $\cos\theta$ 和 $\sin\theta$, 设 $x=\tan\theta$，由三角函数公式得

$$x^2+2\tau x-1=0. \tag{5.88}$$

x 有两个根，为保证 $|\theta| \le \frac{\pi}{4}$ 取其中绝对值较小的一个，为

$$x = \tan\theta = \mathrm{sgn}(\tau)(-|\tau| + \sqrt{\tau^2+1}) = \frac{\mathrm{sgn}(\tau)}{|\tau| + \sqrt{\tau^2+1}}, \tag{5.89}$$

(其中 sgn(·) 表示符号函数，对非负数取 1，对负数取-1) 从 $x = \tan\theta$ 再计算出

$$\cos\theta = (1+x^2)^{-\frac{1}{2}}, \quad \sin\theta = x\cos\theta. \tag{5.90}$$

这样求 $\cos\theta$ 和 $\sin\theta$ 避免了三角函数计算并且 x 的计算方法考虑到了避免两个相近数相减造成精度损失的问题。

设 A 为实对称阵，J_{ij} 是 A 关于下标 (i,j) 的的 Jacobi 变换阵，令 $A^* = J_{ij}AJ_{ij}^T$，则 A^* 有如下性质:

i) A^*仍为对称阵;

ii) $a^*_{ij} = a^*_{ji} = 0$, 且对$k,t \ne i,j$有$a^*_{kt} = a_{kt}$;

iii) 对$t \ne i,j$有$(a^*_{it})^2 + (a^*_{jt})^2 = a^2_{it} + a^2_{jt}$;

iv) $\sum_i \sum_j (a^*_{ij})^2 = \sum_i \sum_j a^2_{ij}$;

v) $\mathrm{off}(A^*) = \mathrm{off}(A) - 2a^2_{ij}$, 其中 off($A$) 表示 A 中非对角线元素的平方和.

Jacobi 算法从 $A^{(0)} = A$ 和 $U^{(0)} = I_n$ 出发反复作 Jacobi 变换，设已有 $A^{(k-1)}$, 则在第 k 步求 $A^{(k-1)}$ 的非对角元素中绝对值最大者，设为其 (i_k, j_k) 元素，用(5.87)和(5.90)针对 $A^{(k-1)}$ 和 (i_k, j_k) 求出 Jacobi 变换矩阵 $J^{(k)}$, 则令 $U^{(k)} = U^{(k-1)}(J^{(k)})^T$, $A^{(k)} = J^{(k)}A^{(k-1)}(J^{(k)})^T$, 如此重复直到 $A^{(k)}$ 的非对角元素的绝对值最大值小于预定的精度 ϵ。这时有 $A = U\Lambda U^T$, $U = U^{(k)}$ 是正交阵，Λ 近似为对角阵。

可以证明上述 Jacobi 算法的 $A^{(k)}$ 收敛到一个对角阵且对角线元素为 A 的特征值。此算法收敛较快，但每次寻找非对角元素中绝对值最大的一个比较耗时。改进的 Jacobi 算法从 A 和 $U = I$ 出发，在第 k 步时基于上一步的 A 计算一个界限 $\epsilon_k = \sqrt{\mathrm{off}(A^{(k)})}/[n(n-1)]$, 然后对 A 的每个严格上三角元素都做一次 Jacobi 变换，用变换后的矩阵代替原来的 A 并更新矩阵 U，但是若该严格上三角元素绝对值小于 ϵ_k 就跳过该元素。所有严格上三角元素都处理过一遍才进入第 $k+1$ 步并计算新的 ϵ_{k+1}, 重复运算直到 ϵ_{k+1} 小于预先指定的误差限 ϵ 为止。

在 R 软件中，用 `eigen()` 函数计算特征值和特征向量。

5.5.3 用 QR 分解方法求对称矩阵特征值分解 *

计算实对称矩阵特征值分解的一种较好的方法是利用 Householder 变换和 Givens 变换。首先，若 A 为 n 阶实对称矩阵，可以用 $n-2$ 个 Householder 变换把它变成对称三

对角矩阵: 设 H_1 为一个分块对角矩阵，主对角线的第一块为 1 阶单位阵，第二块是把 A 的第一列最后 $n-1$ 个元素中后 $n-2$ 个元素变成零的 Householder 变换阵，则 H_1A 第一列为 $(a_{11}, a_{21}^{(1)}, 0, \ldots, 0)^T$, 其中 $a_{21}^{(1)} = \sqrt{a_{21}^2 + \cdots + a_{n1}^2}$，且 H_1A 的第一行与 A 的第一行完全相同，于是，H_1AH_1 的第一行为 $(a_{11}, a_{21}^{(1)}, 0, \ldots, 0)^T$, 注意到 H_1 的对称性与正交性，H_1AH_1 仍为对称阵，但是第一列和第一行的最后 $n-2$ 个元素已经变成了零。在第二步，可以构造一个分块对角矩阵 H_2，对角线第一块为 I_2，第二块是把矩阵 H_1AH_1 的第二列中最后 $n-3$ 个元素变成零的 Householder 变换矩阵，把 H_1AH_1 变成 $H_2H_1AH_1H_2$, 易见第一列和第一行不变，第二列和第二行的最后 $n-3$ 个元素变成了零。如此进行下去得到 $A^{(0)} = H_{n-2}\cdots H_1AH_1\cdots H_{n-2}$，使得 $A^{(0)}$ 为三对角对称矩阵。

得到三对角对称矩阵 $A^{(0)}$ 以后，进行 QR 迭代。设经过 $k-1$ 次迭代后得到矩阵 $A^{(k-1)}$，在第 k 步，先选一个平移量 t_k，对矩阵 $A^{(k-1)} - t_kI$ 用 Givens 变换方法作 QR 分解得到 $A^{(k-1)} - t_kI = Q_kR_k$, 把得到的上三角阵 R_k 右乘 Q_k 再反向平移，得到

$$A^{(k)} = R_kQ_k + t_kI = Q_k^T(A^{(k-1)} - t_kI)Q_k + t_kI = Q_k^TA^{(k-1)}Q_k, \tag{5.91}$$

这样的 $A^{(k)}$ 仍是三对角对称矩阵，如此迭代直到 $A^{(k)}$ 变成对角形。收敛时，$A^{(k)}$ 的对角线元素为各个特征值，$H_1\cdots H_{n-2}Q_1\cdots Q_k$ 的各列为相应特征向量。

具体的算法比较复杂，详见 Monahan(2001)[31] §6.5, Gentle(2007)[19] §7.4。

5.5.4 奇异值分解的计算 *

设 A 为任意非零 $n\times m$ 实值矩阵，先说明 A 的奇异值分解的存在性。以下用 $\|\cdot\|$ 表示向量的长度，即 $\|\cdot\|_2$。首先，非负定阵 A^TA 和 AA^T 有共同的正特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ 且个数 r 为矩阵 A 的秩。设 $\boldsymbol{u}^{(i)}$ 是 A^TA 的特征值 λ_i 对应的单位特征向量（长度为 1），则 $A^TA\boldsymbol{u}^{(i)} = \lambda_i\boldsymbol{u}^{(i)}$, 由此式可得 $\|A\boldsymbol{u}^{(i)}\|^2 = \lambda_i$。在 $A^TA\boldsymbol{u}^{(i)} = \lambda_i\boldsymbol{u}^{(i)}$ 两边左乘矩阵 A 得到 $(AA^T)(A\boldsymbol{u}^{(i)}) = \lambda_i(A\boldsymbol{u}^{(i)})$, 即 $A\boldsymbol{u}^{(i)}$ 是矩阵 AA^T 的对应于特征值 λ_i 的特征向量，长度为 $d_i = \sqrt{\lambda_i}$。设 AA^T 的所有特征向量组成的正交阵为 V, $\boldsymbol{v}^{(j)}$ 是 V 的第 j 列，适当构造的 V 可使得

$$\boldsymbol{v}^{(j)}A\boldsymbol{u}^{(i)} = d_i\delta_{i-j},\ i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n, \tag{5.92}$$

其中 δ_{i-j} 是 Kronecker 记号，当 $i = j$ 时表示 1，当 $i \neq j$ 时表示 0, 当 $i > r$ 时令 $d_i = 0$。把(5.92)式写成矩阵形式即 $V^TAU = D$，$A = VDU^T$(D 的定义见(5.83))，说明任何非零矩阵 A 均有奇异值分解。

以上的证明给出了求奇异值分解的一种方法：先选 A^TA 和 AA^T 中阶数较低一个, 不妨设是 A^TA，求其特征值分解得到 A 的所有奇异值和矩阵 U，然后利用上面的关系得到 AA^T

的对应于非零特征值的特征向量，如果需要再补充适当列向量组成正交方阵 V 即可。这种方法比较简单，但是计算 A^TA 会造成累积误差。

当 A 为 $n\times m$ 的列满秩矩阵时，可以用类似 §5.5.3的 QR 分解方法来求 A 的奇异值分解。方法描述如下。仿照 §5.5.3把一个对称矩阵变成三对角矩阵的做法，我们可以先在 A 的左边乘以 $n-2$ 个对称正交阵把 A 变成上三角形，然后在此上三角形矩阵右边乘以 $n-1$ 个对称正交阵将其变成上双对角阵，即除对角线和上副对角线外的元素都是零的矩阵。总之，存在 n 阶正交阵 P 和 m 阶正交阵 Q 使得

$$PAQ=\begin{pmatrix} B_{m\times m} \\ \mathbf{0}_{(n-m)\times m} \end{pmatrix}, \tag{5.93}$$

其中 B 的元素满足当 $i>j$ 或 $j>i+1$ 时 $b_{ij}=0$, 且 B 满秩。

得到上双对角阵 B 后，先求 B^TB 的特征值分解。B^TB 是一个三对角对称阵，可以用 §5.5.3的 QR 方法求特征值分解。设 $B^TB=U_1D_m^2U_1^T$, 其中 $D_m=\mathrm{diag}(\sqrt{\lambda_1},\dots,\sqrt{\lambda_m})$, $\lambda_1\geq\cdots\geq\lambda_m>0$ 为 B^TB 的所有特征值，U_1 各列为 B^TB 的特征向量。令 $V_1=BU_1D_m^{-1}$, 则 $V_1^TV_1=I_m$, 于是 $B=V_1D_mU_1^T$ 是 B 的奇异值分解。

这时，

$$\begin{aligned} PAQ&=\begin{pmatrix} B \\ \mathbf{0} \end{pmatrix}=\begin{pmatrix} V_1D_mU_1^T \\ \mathbf{0} \end{pmatrix}=\begin{pmatrix} V_1 & \mathbf{0} \\ \mathbf{0} & I_{n-m} \end{pmatrix}_{n\times n}\begin{pmatrix} D_m \\ \mathbf{0} \end{pmatrix}_{n\times m}U_1^T \\ &\triangleq V_2\begin{pmatrix} D_m \\ \mathbf{0} \end{pmatrix}_{n\times m}U_1^T, \end{aligned} \tag{5.94}$$

则 V_2 为 n 阶正交阵，可得 A 有奇异值分解

$$A=(P^TV_2)D(QU_1)^T\triangleq VDU^T, \tag{5.95}$$

其中 $V=P^TV_2$ 为 n 阶正交阵，$U=QU_1$ 为 m 阶正交阵，D 为 $n\times m$ 对角阵，对角线元素为 B^TB 的特征值的算数平方根。

5.6 广义逆矩阵

当 A 为满秩方阵时，线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 有唯一解 $\boldsymbol{x}=A^{-1}\boldsymbol{b}$。当 A 为不满秩的方阵或长方形 $n\times m$ 矩阵时，A^{-1} 不存在，这时能否用类似逆矩阵的方式表示线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$ 的解？$A\boldsymbol{x}=\boldsymbol{b}$ 可能有唯一解、无穷多个解或无解（无解时可以找最小二乘解），用广义逆矩阵可以统一地给出这些问题的解。

一些统计计算问题的理论研究和数值计算也用到广义逆矩阵，比如，线性模型中参数最小二乘估计的表示, 典型相关分析中典型相关系数和典型变量的计算。

广义逆有多种定义，最常用的一种是加号逆。

定义 (加号逆) 设 A 为 $n\times m$ 矩阵，若 $m\times n$ 矩阵 G 满足

$$\begin{aligned}&\text{i)}AGA=A;\\&\text{ii)}GAG=G;\\&\text{iii)}(AG)^T=AG;\\&\text{iv)}(GA)^T=GA\end{aligned}$$

则称矩阵 G 为矩阵 A 的**加号逆**或 Moore-Penrose 广义逆，记作 A^+。如果 G 满足定义中的第一个条件，则称 G 为 A 的**减号逆**，记作 A^-。显然，如果 A 本身就是 n 阶可逆方阵，则 A^{-1} 满足上述四个条件。

定理 5.6.1. $n\times m$ 矩阵 A 的加号逆存在且唯一。

证明: 若 A 不是零矩阵（所有元素都等于零的矩阵），则 A 有奇异值分解 $A=VDU^T$，取 $G=UD^+V^T$, 其中 D^+ 是把对角阵 D 的主对角线中非零元素换成相应元素的倒数，其它元素保持为零，容易验证 G 是 A 的加号逆（见习题28）。当 A 为零矩阵时，$m\times n$ 的零矩阵是 A 的加号逆。

若 A_1^+ 和 A_2^+ 是 $n\times m$ 矩阵 A 的两个加号逆，则

$$\begin{aligned}&A_1^+=A_1^+AA_1^+=A_1^+(AA_1^+)^T=A_1^+(A_1^+)^T(A)^T=A_1^+(A_1^+)^T(AA_2^+A)^T\\=&A_1^+(A_1^+)^TA^T(AA_2^+)^T=A_1^+(AA_1^+)^TAA_2^+=A_1^+AA_1^+AA_2^+=A_1^+AA_2^+\\=&(A_1^+A)^TA_2^+=A^T(A_1^+)^TA_2^+=(AA_2^+A)^T(A_1^+)^TA_2^+=(A_2^+A)^TA^T(A_1^+)^TA_2^+\\=&(A_2^+A)^T(A_1^+A)^TA_2^+=A_2^+AA_1^+AA_2^+=A_2^+AA_2^+=A_2^+\end{aligned}$$

可见加号逆存在唯一。 □

以上定理证明中给出了用奇异值分解计算加号逆的方法。另一种计算加号逆的方法是利用“满秩分解”。设 A 为非零的 $n\times m$ 实矩阵，设 $\mathrm{rank}(A)=r$, 则存在 $n\times r$ 的满秩矩阵 B 和 $r\times m$ 的满秩矩阵 C 使得 $A=BC$, 这称为 A 的**满秩分解** (见习题30)。这时，A 的加号逆可表示为 (见习题31)

$$A^+=C^T(CC^T)^{-1}(B^TB)^{-1}B^T. \tag{5.96}$$

加号逆也是减号逆，所以减号逆一定存在，但是当 A 不是满秩方阵时 A 的减号逆有无穷多个。

利用广义逆可以讨论线性方程组解的表示。

定理 5.6.2. 系数矩阵为 $n \times m$ 矩阵的线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 有解的充分必要条件为

$$AA^{+}\boldsymbol{b} = \boldsymbol{b}. \tag{5.97}$$

在方程组有解时，对 A 的任何一个减号逆 A^{-}, $\boldsymbol{x} = A^{-}\boldsymbol{b}$ 是方程组的解，且方程组的通解为

$$\boldsymbol{x} = A^{+}\boldsymbol{b} + (I - A^{+}A)\boldsymbol{y},\ \forall \boldsymbol{y} \in \mathbb{R}^{n}. \tag{5.98}$$

证明: 若(5.97)成立，则 $\boldsymbol{x} = A^{+}\boldsymbol{b}$ 是一个解。反之，若 $\boldsymbol{x}$ 满足 $A\boldsymbol{x} = \boldsymbol{b}$, 则

$$\boldsymbol{b} = A\boldsymbol{x} = AA^{+}A\boldsymbol{x} = AA^{+}\boldsymbol{b}$$

成立，即(5.97)成立。

设 A^{-} 是 A 的任意一个减号逆，若 $\boldsymbol{x} = A^{-}\boldsymbol{b}$，则

$$\begin{aligned} A\boldsymbol{x} =& AA^{-}\boldsymbol{b} = AA^{-}(AA^{+}\boldsymbol{b}) \quad (\text{由(5.97)式}) \\ =& AA^{+}\boldsymbol{b} = \boldsymbol{b} \quad (\text{由(5.97)式}). \end{aligned}$$

令 $\boldsymbol{x}$ 由(5.98)定义，容易验证它是一个解。反之，若 $\boldsymbol{x}$ 满足 $A\boldsymbol{x} = \boldsymbol{b}$，则

$$\boldsymbol{x} = A^{+}\boldsymbol{b} + \boldsymbol{x} - A^{+}\boldsymbol{b} = A^{+}\boldsymbol{b} + \boldsymbol{x} - A^{+}A\boldsymbol{x} = A^{+}\boldsymbol{b} + (I - A^{+}A)\boldsymbol{x},$$

满足通解形式。实际上，通解(5.98)中的 A^{+} 也可以替换成 A 的任何一个减号逆。 □

当线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 有解时，$\boldsymbol{x} = A^{+}\boldsymbol{b}$ 是所有解中唯一的长度最小的解（参见习题35）。

容易看出，当且仅当 A 为满秩方阵时 $A^{+} = A^{-1}$。当且仅当 A 为 $n \times m$ 列满秩阵时 $A^{+}A = I_m$, 这时 $A^{+} = (A^{T}A)^{-1}A^{T}$，当且仅当 A 为 $n \times m$ 行满秩阵时 $AA^{+} = I_n$，这时 $A^{+} = A^{T}(AA^{T})^{-1}$。

定理 5.6.3. 加号逆有如下的性质:

i) $(A^{+})^{+} = A$;

ii) $(A^{T})^{+} = (A^{+})^{T}$;

iii) $(\lambda A)^{+} = \lambda^{-1}A^{+},\ \forall \lambda \neq 0$;

iv) $\mathrm{rank}(A^{+}) = \mathrm{rank}(A) = \mathrm{rank}(AA^{+}) = \mathrm{rank}(A^{+}A)$;

v) $(A^{T}A)^{+} = A^{+}(A^{+})^{T}$;

vi) $A^{+} = (A^{T}A)^{+}A^{T} = A^{T}(AA^{T})^{+}$;

vii) AA^{+}和$A^{+}A$都是对称幂等矩阵;

viii) 若 A 是对称幂等矩阵，则 $A^{+} = A$。

证明留给读者 (习题33)。

广义逆可以用来分析回归分析和线性模型问题中最小二乘解的结构。设 X 为 $n\times m$ 矩阵 $(n>m)$，则当 X 列满秩时矩阵 $P=X(X^TX)^{-1}X^T$ 是对称幂等矩阵，可以把向量 $\boldsymbol{y}$ 正交投影到 X 的各列张成的线性空间 $\mu(X)$ 中, 这时最小二乘问题

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m}\|\boldsymbol{y}-X\boldsymbol{\beta}\|_2^2 \tag{5.99}$$

有唯一解 $\hat{\boldsymbol{\beta}}=(X^TX)^{-1}X^T\boldsymbol{y}$。对一般情况有如下结论。

定理 5.6.4. 设 X 为 $n\times m$ 矩阵 $(n>m)$, 则最小二乘问题(5.99)的所有的最小二乘解可以写成

$$\hat{\boldsymbol{\beta}}=X^+\boldsymbol{y}+(I-X^+X)\boldsymbol{z},\ \forall\boldsymbol{z}\in\mathbb{R}^m. \tag{5.100}$$

在这些最小二乘解中 $\boldsymbol{\beta}_0=X^+\boldsymbol{y}$ 是唯一的长度最短的解。

证明: 令 $P=XX^+$，则 P 是对称幂等矩阵，$P\boldsymbol{y}$ 是向量 $\boldsymbol{y}$ 到 X 的各列张成的线性子空间 $\mu(X)$ 的正交投影 (见习题34), 且

$$X\hat{\boldsymbol{\beta}}=XX^+\boldsymbol{y}+(X-XX^+X)\boldsymbol{z}=XX^+\boldsymbol{y}=P\boldsymbol{y},$$

于是

$$\begin{aligned}\|\boldsymbol{y}-X\boldsymbol{\beta}\|_2^2=&\|\boldsymbol{y}-X\hat{\boldsymbol{\beta}}+X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\\=&\|\boldsymbol{y}-P\boldsymbol{y}+X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\\=&\|\boldsymbol{y}-P\boldsymbol{y}\|^2+\|X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\quad(\text{因为 }P\boldsymbol{y}\text{ 是正交投影})\\=&\|\boldsymbol{y}-X\hat{\boldsymbol{\beta}}\|^2+\|X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2\\\geq&\|\boldsymbol{y}-X\hat{\boldsymbol{\beta}}\|^2,\end{aligned}$$

所以(5.100)是最小二乘问题(5.99)的解, 等号成立当且仅当 $X(X^+y-\boldsymbol{\beta})=0$。

若 $\tilde{\boldsymbol{\beta}}$ 是最小二乘问题(5.99)的解，则 $X(X^+\boldsymbol{y}-\tilde{\boldsymbol{\beta}})=0$, 于是 $\tilde{\boldsymbol{\beta}}$ 是线性方程组 $X\tilde{\boldsymbol{\beta}}=XX^+\boldsymbol{y}$ 的解，由定理5.6.2可知存在 $\boldsymbol{z}$ 使得

$$\tilde{\boldsymbol{\beta}}=X^+(XX^+\boldsymbol{y})+(I-X^+X)\boldsymbol{z}=X^+\boldsymbol{y}+(I-X^+X)\boldsymbol{z}.$$

设 $\hat{\boldsymbol{\beta}}$ 为最小二乘解(5.100)，则

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}\|_2^2=&\|X^+\boldsymbol{y}\|^2+\|(I-X^+X)\boldsymbol{z}\|^2+2\boldsymbol{z}^T(I-X^+X)X^+\boldsymbol{y}\\=&\|X^+\boldsymbol{y}\|^2+\|(I-X^+X)\boldsymbol{z}\|^2\\\geq&\|X^+\boldsymbol{y}\|^2,\end{aligned}$$

等号成立当且仅当 $(I-X^+X)\boldsymbol{z}=0$ 即 $\hat{\boldsymbol{\beta}}=X^+\boldsymbol{y}$。 □

习题五

1. 设 A 为 $n\times m$ 矩阵，$\boldsymbol{x}$ 为 m 维向量，为了计算 $A\boldsymbol{x}$，可以逐个计算结果的 n 个元素，也可以把这个乘法看成是对 A 的各列用 $\boldsymbol{x}$ 作为线性组合系数作线性组合，逐次把 A 的第 j 列与 x_j 相乘累加到结果向量中。写出这两种算法，尽可能减少不必要的存储。计算这两种算法所需的乘法和加法的次数。如果使用 C 语言或 FORTRAN 语言来实现这两种算法，在 n,m 很大时，两种算法会有不同的计算效率，设法验证。

2. 对满秩 n 阶上三角矩阵 A，编写算法求解线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$, 要求把 $\boldsymbol{x}$ 的解保存在 $\boldsymbol{b}$ 原来的存储空间中。计算算法需要的乘除法和加减法次数。

3. 对 n 阶单位下三角矩阵 A，编写算法求解线性方程组 $A\boldsymbol{x}=\boldsymbol{b}$, 要求把 $\boldsymbol{x}$ 的解保存在 $\boldsymbol{b}$ 原来的存储空间中。计算算法需要的乘除法和加减法次数。

4. 设 A 为上三角矩阵，$\boldsymbol{e}_k=(0,\dots,0,1,0,\dots,0)^T$ 为单位向量，写算法求解 $A\boldsymbol{x}=\boldsymbol{e}_k$。解 $\boldsymbol{x}$ 中那些元素是一定等于零的？如何简化求解过程？编写算法求 A^{-1} 并把 A^{-1} 保存在 A 原来的存储空间中。

5. 证明5.2.2中用列主元法进行矩阵 LU 分解(5.18)的算法正确。

   (1) 对 $M^{(i)}=I_n-\boldsymbol{m}^{(i)}\boldsymbol{e}_i^T$ 矩阵，证明当 $k,j>i$ 时有

   $$P(k,j)M^{(i)}=[I_n-P(k,j)\boldsymbol{m}^{(i)}\boldsymbol{e}_i^T]P(k,j).$$

   (2) 令 $M_*^{(k)}=I_n-\boldsymbol{m}_*^{(k)}\boldsymbol{e}_k^T$ ($\boldsymbol{m}_*^{(k)}$ 定义见(5.20))，证明

   $$\begin{aligned}&M^{(n-1)}P(n-1,s_{n-1})\dots M^{(2)}P(2,s_2)M^{(1)}P(1,s_1)\\=&M_*^{(n-1)}\dots M_*^{(1)}P(n-1,s_{n-1})\dots P(1,s_1).\end{aligned}$$

   (3) 证明

   $$(M_*^{(k+1)}M_*^{(k)})^{-1}=I_n+\boldsymbol{m}_*^{(k)}\boldsymbol{e}_k^T+\boldsymbol{m}_*^{(k+1)}\boldsymbol{e}_{k+1}^T.$$

   (4) 证明 $PA=LU$ 成立。

6. 用 R 语言编写用列主元法作高斯消元同时得到矩阵 LU 分解的算法程序。

7. 编写 Cholesky 分解算法的程序。对输入的正定阵 A（只有下三角部分需要输入值）, 求 Cholesky 分解 $A=LL^T$ 并把矩阵 L 存放在 A 的下三角部分中作为函数返回值。(如果使用 R 语言的编写一个输入 A 输出 L 的函数，由于 R 语言的设计特点，这样修改的是矩阵 A 的一个副本)。

8. 编写 Cholesky 分解算法的程序。设正定阵 A 只保存了下三角部分，元素按行次序保存在一个一维数组中输入，Cholesky 分解 $A = LL^T$ 得到的下三角矩阵 L 保存到 A 所用的存储空间中返回。

9. 对正定阵 A，证明(5.29)定义的向量与 A 组成的二次型 $\boldsymbol{\alpha}^T A\boldsymbol{\alpha}$ 等于(5.27)的右边。

10. 设 A 为正定阵，给出计算 $\boldsymbol{x}^T A^{-1}\boldsymbol{y}$ 的有效算法。

11. 考虑线性回归模型 $\boldsymbol{y} = X\boldsymbol{\beta}+\boldsymbol{\varepsilon}$，其中 X 为 $n\times p$ 矩阵 $(n > p)$, 设 X 列满秩，则回归系数 $\boldsymbol{\beta}$ 的最小二乘估计为 $\hat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\boldsymbol{y}$, 残差平方和 SSE 为 $\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{y}^T X(X^T X)^{-1}X^T\boldsymbol{y}$。写出利用 Cholesky 分解方法计算 $\hat{\boldsymbol{\beta}}$, SSE 和 $(X^T X)^{-1}$ 的算法。

12. 证明矩阵范数的三个公式(5.37)–(5.39)。

13. 设方阵 A 满秩，$a_{ij} = 0$ 对 $i > j + p$ 和 $j > i + q$, 若 A 有 LU 分解 $A = LU$，证明当 $i > j + p$ 时 $l_{ij} = 0$, 当 $j > i + q$ 时 $u_{ij} = 0$。

14. 证明5.3.1关于三对角矩阵的递推算法。

15. 编写用 LU 分解方法求解三对角矩阵的算法程序。要在程序中设置分母等于零时快速算法失败的预防措施。

16. 对例5.3.1, 编写输入 b 和 δ 计算对数似然函数(5.48)的高效算法程序, 尽可能不占用额外存储空间。

17. 编写算法程序,输入时间序列自协方差函数值 $(\gamma(0), \gamma(1), \ldots, \gamma(n))$ 和实数值 $(y_1, y_2, \ldots, y_n)$，用递推算法求解方程(5.50)和(5.51), 输出 $\boldsymbol{a}^{(n)}$ 和 $\boldsymbol{x}^{(n)}$。注意避免中间结果不必要的存储。

18. 对 $n\times p$ 矩阵 X（设 $n > p$）编写用 Gram-Schmidt 正交化方法和修正的 Gram-Schmidt 方法作 QR 分解的 R 程序。

19. 对 $n \times p$ 矩阵 X（设 $n > p$）编写用 Householder 变换方法作 QR 分解的 R 程序。

20. 对 $n \times p$ 矩阵 X（设 $n > p$）编写用 Givens 变换方法作 QR 分解的 R 程序。

21. 对线性模型 (5.64), 设 X 列满秩，编写用 QR 分解求 $\hat{\boldsymbol{\beta}}$、SSE、回归残差 $\boldsymbol{e} = \boldsymbol{y} - X\boldsymbol{\beta}$ 和 $(X^T X)^{-1}$ 的 R 程序。

22. 编写关于实对称矩阵 A 用 Jacobi 方法求特征值分解的 R 程序。

23. 编写关于实对称矩阵 A 用改进的 Jacobi 方法求特征值分解的 R 程序。

24. 设 A 为 n 阶实对称方阵，编写 R 语言程序用正交相似变换 $B = QAQ$ 把 A 变换为三对角对称矩阵 B，其中 Q 为 n 阶对称正交阵, 输出 B 和 Q 的值。

25. 设 n 方阵 A 为上双对角矩阵: 当 $j \neq i, i+1$ 时总有 $a_{ij} = 0$。证明 $B = A^TA$ 为三对角矩阵。

26. 设 A 为 n 阶非负定对称矩阵，用拉格朗日乘子法求 $\boldsymbol{x}^T\boldsymbol{x} = 1$ 条件下 $\boldsymbol{x}^TA\boldsymbol{x}$ 的最大值点。

27. 设 A 为 n 阶非负定对称矩阵，用 A 的特征值分解求 $\boldsymbol{x}^T\boldsymbol{x} = 1$ 条件下 $\boldsymbol{x}^TA\boldsymbol{x}$ 的最大值点。

28. 设 $n \times m$ 非零矩阵 A 的有奇异值分解 $A = VDU^T$，证明 UD^+V^T 为其加号逆, 其中 D^+ 是把对角阵 D 的主对角线中非零元素换成相应元素的倒数，其它元素保持为零。

29. 设 A 为 n 阶实对称方阵，B 为 n 阶正定阵。写出用 Cholesky 分解的方法求解广义特征值问题 $A\boldsymbol{\alpha} = \lambda B\boldsymbol{\alpha}$ 的算法，并用编写 R 程序实现该算法。

30. 设 A 为非零的 $n \times m$ 实矩阵，且 $\mathrm{rank}(A) = r$, 证明存在 $n \times r$ 的满秩矩阵 B 和 $r \times m$ 的满秩矩阵 C 使得 $A = BC$。

31. 设 $n \times m$ 非零的实矩阵 A 有满秩分解 $A = BC$, 证明 A 的加号逆可表示为

$$A^+ = C^T(CC^T)^{-1}(B^TB)^{-1}B^T.$$

32. 设 X 为 $n \times p$ 矩阵（$n > p$），$\boldsymbol{y}$ 为 n 维向量，证明正规方程 $X^TX\boldsymbol{\beta} = X^T\boldsymbol{y}$ 的解 $\boldsymbol{\beta}$ 中长度最小的一个为 $X^+\boldsymbol{y}$。

33. 证明加号逆的性质 i)—viii)。

34. 对 $n \times m$ 矩阵 A，令 $P = A(A^TA)^+A^T$, 证明 $P = AA^+$, 且 P 是对称幂等阵。记 $\mu(A) = \{A\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^m\}$, 证明对任意 $\boldsymbol{y} \in \mathbb{R}^n$ 有 $P\boldsymbol{y} \in \mu(A)$ 且

$$\boldsymbol{z}^T(\boldsymbol{y} - P\boldsymbol{y}) = 0, \ \forall \boldsymbol{z} \in \mu(A).$$

35. 设 A 为 $n \times m$ 矩阵，若线性方程组 $A\boldsymbol{x} = \boldsymbol{b}$ 有解，证明 $\boldsymbol{x}_0 = A^+\boldsymbol{b}$ 是所有解中唯一的长度最小的解。

第六章　最优化与方程求根

6.1　最优化问题和求解

在统计计算中广泛用到求最小（最大）值点或求方程的根的算法，比如，参数最大似然估计、置信区间的统计量法、回归分析参数估计、惩罚似然估计、惩罚最小二乘估计，等等。求最小值点或最大值点的问题称为**最优化**问题（或称优化问题），最优化问题和求方程的根经常具有类似的算法，所以在这一章一起讨论。有些情况下，可以得到解的解析表达式；更多的情况只能通过数值迭代算法求解。

本章将叙述最小值点存在性的一些结论，并给出一些常用的数值算法。还有很多的优化问题需要更专门化的算法。关于最优化问题的详细理论以及更多的算法，读者可以参考这方面的教材和专著，如徐成贤等 (2002)[10]、高立 (2014)[3]、Lange(2013)[25] 。

因为求最大值的问题和求最小值的问题完全类似，而且最小值问题涉及到凸函数、正定海色阵等，比最大值问题方便讨论，所以我们只考虑最小值问题。

设 $f(\boldsymbol{x})$ 是 $\mathbb{R}^d$ 上的多元函数，如果 $f(\boldsymbol{x})$ 有一阶偏导数，则记 $f(\boldsymbol{x})$ 的梯度向量为

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \frac{\partial f(\boldsymbol{x})}{\partial x_2}, \ldots, \frac{\partial f(\boldsymbol{x})}{\partial x_d} \right)^T, \tag{6.1}$$

有的教材记梯度向量为 $\boldsymbol{g}(\boldsymbol{x})$。如果 $f(\boldsymbol{x})$ 的二阶偏导数都存在，则记它的海色阵为

$$\nabla^2 f(\boldsymbol{x}) = \left(\frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j} \right)_{\substack{i=1,\ldots,d \\ j=1,\ldots,d}}, \tag{6.2}$$

有的教材记海色阵为 $H_f(\boldsymbol{x})$ 或 $H(x)$。当所有二阶偏导数连续时，海色阵是对称阵。

6.1.1 优化问题的类型

定义 (无约束最优化) 设 $f(\boldsymbol{x}), \boldsymbol{x}=(x_1,\ldots,x_d)^T \in \mathbb{R}^d$ 是实数值的 d 元函数，求

$$\arg\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \tag{6.3}$$

的问题称为（全局的）**无约束最优化**问题。这里，符号 $\arg\min$ 表示求最小值点的运算, 称 $f(\boldsymbol{x})$ 为目标函数。$\boldsymbol{x}^*$ 是问题的解等价于

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*),\ \forall \boldsymbol{x} \in \mathbb{R}^d,$$

称这样的 $\boldsymbol{x}^*$ 为 $f(\boldsymbol{x})$ 的一个**全局最小值点**。全局最小值点不一定存在，存在时不一定唯一。如果

$$f(\boldsymbol{x}) > f(\boldsymbol{x}^*),\ \forall \boldsymbol{x} \in \mathbb{R}^d \backslash \{\boldsymbol{x}^*\},$$

则称 $\boldsymbol{x}^*$ 为 $f(\boldsymbol{x})$ 的一个**全局严格最小值点**, 全局严格最小值点如果存在一定是唯一一个。上式中 $\backslash$ 表示集合的差。

定义 (约束最优化) 设 $c_i(\cdot),\ i=1,\ldots,p+q$ 是 d 元实值函数, 求

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}),\ \text{s.t.} \\ c_i(\boldsymbol{x}) = 0,\ i=1,\ldots,p \\ c_i(\boldsymbol{x}) \geq 0,\ i=p+1,\ldots,p+q \end{cases} \tag{6.4}$$

的问题称为（全局的）**约束最优化**问题, 这里 s.t.(subject to) 是“满足如下约束条件”的意思, $c_i, i=1,\ldots,p$ 称为**等式约束**，$c_i, i=p+1,\ldots,p+q$ 称为**不等式约束**。满足约束的 $\boldsymbol{x}\in\mathbb{R}^d$ 称为一个**可行点**, 所有可行点的集合 D 称为**可行域**:

$$D = \{\boldsymbol{x}\in\mathbb{R}^d:\ c_i(\boldsymbol{x})=0, i=1,\ldots,p;\ c_i(\boldsymbol{x})\geq 0, i=p+1,\ldots,p+q\}.$$

最优化问题(6.4)也可以写成 $\min\limits_{\boldsymbol{x}\in D} f(\boldsymbol{x})$。若 $\boldsymbol{x}*\in D$ 使得

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*),\ \forall \boldsymbol{x} \in D,$$

称 $\boldsymbol{x}^*$ 为约束优化问题(6.4)的一个**全局最小值点**。如果 $\boldsymbol{x}^*\in D$ 使得

$$f(\boldsymbol{x}) > f(\boldsymbol{x}^*),\ \forall \boldsymbol{x} \in D\backslash\{\boldsymbol{x}^*\},$$

称 $\boldsymbol{x}^*$ 为约束优化问题(6.4)的一个**全局严格最小值点**。

最优化问题经常使用数值迭代方法求解，数值迭代方法一般每一步都得到比上一步更小的目标函数值，这样最后迭代结束时往往得到的不是全局的最小值点，而是局部的极小值点。设 D 为可行域 (无约束问题 $D=\mathbb{R}^d$)，若对点 $\boldsymbol{x}^*$ 存在一个邻域 $U(\boldsymbol{x}^*)=\{\boldsymbol{x}\in\mathbb{R}^d:\|\boldsymbol{x}-\boldsymbol{x}^*\|<\delta\}(\delta>0)$，使得 $f(\boldsymbol{x})\geq f(\boldsymbol{x}^*),\forall\boldsymbol{x}\in U(\boldsymbol{x}^*)\cap D$，称 $\boldsymbol{x}^*$ 是 $f(\cdot)$ 的一个**局部极小值点**。如果 $\boldsymbol{x}^*$ 使得 $f(\boldsymbol{x})>f(\boldsymbol{x}^*),\forall\boldsymbol{x}\in U_0(\boldsymbol{x}^*)\cap D$，则称 $\boldsymbol{x}^*$ 为 $f(\boldsymbol{x})$ 的一个**局部严格极小值点**。其中 $U_0(\boldsymbol{x}^*)=\{\boldsymbol{x}\in\mathbb{R}^d:0<\|\boldsymbol{x}-\boldsymbol{x}^*\|<\delta\}$ 表示空心邻域。要注意的是，全局的最小值点一定也是局部极小值点, 另外，极小值点有时不存在，存在时可以不唯一。本书中的最优化算法一般只能收敛到局部极小值点。

例 6.1.1. 函数 $f(x)=\frac{1}{1+x^2},x\in\mathbb{R}$ 没有极小值点，$f(x)>0,\lim_{x\to\pm\infty}f(x)=0$。 □

例 6.1.2. 函数 $f(x)=\max(|x|,1)$ 最小值为 1, 且当 $x\in[-1,1]$ 时都取到最小值。 □

在最优化问题中如果要求解 $\boldsymbol{x}$ 的所有或一部分分量取整数值，这样的问题称为**整数规划**问题。如果要同时考虑具有相同自变量的多个目标函数的优化问题，则将其称为**多目标规划**问题。对约束最优化问题(6.4)，如果 $f(\boldsymbol{x}),c_i(\boldsymbol{x})$ 都是线性函数，称这样的问题为**线性规划**问题。线性规划是运筹学的一个重要工具，在工业生产、经济计划等领域有广泛的应用。如果(6.4)中的 $f(\boldsymbol{x}),c_i(\boldsymbol{x})$ 不都是线性函数，称这样的问题为**非线性规划问题**。对非线性规划问题，如果 $f(\boldsymbol{x})$ 是二次多项式函数，$c_i(\boldsymbol{x})$ 都是线性函数，称这样的问题为**二次规划**问题。本书主要讨论在统计学中常见的优化问题，不讨论线性规划、整数规划、多目标规划等问题。

例 6.1.3. 如下的最优化问题是一个线性规划问题。

$$\begin{cases}\underset{(x_1,x_2)\in\mathbb{R}^2}{\arg\min}\ \ 3x_1-x_2,\ \text{s.t.}\\ x_1+x_2=10,\\ x_1+2x_2+x_3=25,\\ x_1\geq 0,\ x_2\geq 0,x_3\geq 0\end{cases}$$

□

6.1.2 一元函数的极值

先复习一些数学分析中的结论。设 $f(x)$ 为定义在闭区间 $[a,b]$ 上的连续函数，则 $f(x)$ 在 $[a,b]$ 内有界，且能达到最小值和最大值, 极值可以在区间内部或边界取到。如果 $f(x)$ 在区间 (a,b)（有限或无限区间）可微且 $x^*\in(a,b)$ 是 $f(x)$ 的一个局部极小值点，则 $f'(x^*)=0$。另一方面，$f'(x^*)=0$ 不能保证 x^* 是全局或局部的极值点，满足 $f'(x^*)=0$ 的 x^* 叫做 $f(x)$ 的

一个**稳定点**。若 $f'(x^*)=0$ 并且 $f''(x^*)$ 存在，则 $f''(x^*)>0$ 保证了 x^* 是一个局部极小值点，$f''(x^*)<0$ 保证了 x^* 是一个局部极大值点，当 $f''(x^*)=0$ 时则不能确定 x^* 是否极值点。

当一元连续函数 $f(x)$ 的定义域为区间但不是闭区间时，$f(x)$ 在定义域内不一定取到最小值，通常可以通过检查边界处的极限并与内部的所有极小值的比较来确定。

例 6.1.4. 对 $f_1(x)=x^2, x\in\mathbb{R}$, $x^*=0$ 使得 $f_1'(x^*)=2x^*=0$, $f_1''(x^*)=2>0$，是一个局部极小值点。因为 $x^*=0$ 是唯一的极小值点，当 $x\to\pm\infty$ 时函数值增大，所以这个局部极小值点也是全局最小值点。

对 $f_2(x)=x^3$, $x^*=0$ 使得 $f_2'(x^*)=3(x^*)^2=0$, 是一个稳定点。但是 $f_2''(x^*)=6x^*=0$，不能保证 x^* 是极值点。事实上，$x^*=0$ 不是 $f_2(x)=x^3$ 的极值点。

对 $f_3(x)=\frac{1-x}{1+x^2}, x\geq 0$, 有 $f_3(0)=1$, $\lim_{x\to+\infty}f_3(x)=0$, $f_3'(x)=\frac{x^2-2x-1}{(x^2+1)^2}$, 令 $f_3'(x)=0$ 得稳定点 $x^*=1+\sqrt{2}$，且 $f_3(x^*)=-\sqrt{2}/(4+2\sqrt{2})<0$，所以 x^* 是 $f_3(x)$ 的全局最小值点。 □

例 6.1.5. 存在不可微点时极值点也不一定出现在稳定点。例如，设总体 $X\sim \mathrm{U}(0,b)$, 样本 $X_1,\ldots,X_n$, 似然函数为

$$L(b)=\prod_{i=1}^{n}\frac{1}{b}I_{[0,b]}(X_i)=\frac{1}{b^n}I_{[0,b]}(X_{(n)}),$$

其中 $X_{(n)}=\max(X_1,\ldots,X_n)$。导数

$$L'(b)=\begin{cases}0 & b<X_{(n)},\\ \text{不存在}, & b=X_{(n)},\\ -nb^{-n-1}, & b>X_{(n)},\end{cases}$$

$L(b)$ 的最大值点为 $\hat{b}=X_{(n)}$，$L'(\hat{b})$ 不存在。 □

6.1.3 凸函数 *

定义 (**凸集**) 设 S 是 $\mathbb{R}^d$ 的子集，如果 $\forall \boldsymbol{x},\boldsymbol{y}\in S$, $\alpha\in[0,1]$, 都有 $\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}\in S$，则称 S 为**凸集**。

等价地，若对任意正整数 $n\geq 2$ 和 S 中的 n 个互不相同的点 $\boldsymbol{x}^{(j)}, j=1,2,\ldots,n$，以及任意的 $\{\alpha_j, j=1,\ldots,n\}$ 满足 $\alpha_j\geq 0, \sum_{j=1}^n\alpha_j=1$ 都有 $\sum_{j=1}^n\alpha_j\boldsymbol{x}^{(j)}\in S$，则 S 是凸集。

S 是凸集当且仅当 S 中任意两个点的连线都属于 S。例如，实数轴上的凸集和区间是等价的; 平面上边界为椭圆、三角形、平行四边形的集合是凸集, 圆环则不是凸集; 球体、长方体是凸集。

凸集有如下性质:

(1) 凸集的闭包是凸集。

(2) 凸集的内核（所有内点组成的集合）是凸集。

(3) 凸集是连通集合。

(4) 任意多个凸集的交集是凸集。

(5) 关于仿射变换 $f(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$($A$ 为矩阵, $\boldsymbol{b}$ 为常数向量)，凸集的像和原像还是凸集。

(6) 两个凸集 S, T 的笛卡尔积 $S \times T = \{(\boldsymbol{x}, \boldsymbol{y}) : \boldsymbol{x} \in S, \boldsymbol{y} \in T\}$ 是凸集。

(7) 设 S 是 $\mathbb{R}^d$ 中的凸集，λ 为实数，则 $\lambda S \triangleq \{\boldsymbol{y} \in \mathbb{R}^d : \boldsymbol{y} = \lambda \boldsymbol{x}, \boldsymbol{x} \in S\}$ 是凸集。

(8) 设 S, T 是 $\mathbb{R}^d$ 中的两个凸集，则 $S + T \triangleq \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{x} \in S, \boldsymbol{y} \in T\}$ 是凸集。

(9) 对 $\mathbb{R}^d$ 的凸集 S 以及任意 $\boldsymbol{y} \in \mathbb{R}^d$, 令 $\boldsymbol{y}$ 到 S 的距离为 $\operatorname{dist}(\boldsymbol{y}, S) \triangleq \inf_{\boldsymbol{x} \in S} \|\boldsymbol{y} - \boldsymbol{x}\|$, 则至多有一个 $\boldsymbol{x}_0 \in S$ 可以满足 $\|\boldsymbol{y} - \boldsymbol{x}_0\| = \operatorname{dist}(\boldsymbol{y}, S)$。当 S 是闭的凸集时，存在唯一的 $\boldsymbol{x}_0 \in S$ 使得 $\|\boldsymbol{y} - \boldsymbol{x}_0\| = \operatorname{dist}(\boldsymbol{y}, S)$。

(10) 设 S 为凸集，$\boldsymbol{x}_0$ 是 S 的边界点，则存在单位向量 $\boldsymbol{v} \in \mathbb{R}^d$ 使得 $\boldsymbol{v}^T \boldsymbol{x}_0 \geq \boldsymbol{v}^T \boldsymbol{x}, \forall \boldsymbol{x} \in S$。

定义 (**凸函数**) 定义在凸集 S 上的 d 元函数 $f(\boldsymbol{x})$ 称为**凸函数**, 如果对任意 $\boldsymbol{x}, \boldsymbol{y} \in S$ 和 $\alpha \in [0, 1]$ 都有

$$f(\alpha \boldsymbol{x} + (1 - \alpha)\boldsymbol{y}) \leq \alpha f(\boldsymbol{x}) + (1 - \alpha)\boldsymbol{y}, \tag{6.5}$$

如果(6.5)对任意 $\boldsymbol{x} \neq \boldsymbol{y} \in S$ 和 $\alpha \in (0, 1)$ 都满足严格不等式，则称 $f(\boldsymbol{x})$ 为**严格凸函数**。

若 $-f(\boldsymbol{x})$ 是定义在凸集 S 上的凸函数，则称 $f(\boldsymbol{x})$ 为**凹函数**。

对凸函数 $f(\boldsymbol{x})$，设 $\boldsymbol{x}^{(j)}, j = 1, \ldots, n$ 是 S 中的 n 个点，$\alpha_j \geq 0, j = 1, \ldots, n$, $\sum_{j=1}^n \alpha_j = 1$，则

$$f\left(\sum_{j=1}^n \alpha_j \boldsymbol{x}^{(j)}\right) \leq \sum_{j=1}^n \alpha_j f(\boldsymbol{x}^{(j)}).$$

凸函数在最优化问题中有重要作用，原因是, 若 $f(\boldsymbol{x})$ 是凸集 $S \subset \mathbb{R}^d$ 上的凸函数, 那么：

(1) 当 $f(\boldsymbol{x})$ 有一个局部极小值点 $\boldsymbol{x}^*$ 时，这个极小值点也是全局最小值点，但不一定是唯一的全局最小值点，这时集合 $\{\boldsymbol{x} \in S : \ f(\boldsymbol{x}) = f(\boldsymbol{x}^*)\}$ 是一个凸集。

(2) 如果 $\boldsymbol{x}^*$ 是 $f(\boldsymbol{x})$ 的一个稳定点，则 $\boldsymbol{x}^*$ 是 $f(\boldsymbol{x})$ 的一个全局最小值点。

(3) 如果 $f(\boldsymbol{x})$ 是严格凸函数而且全局最小值点存在，这个全局最小值点是唯一的。

对于约束优化问题(6.4), 如果等式约束 $c_i, i = 1, \ldots, p$ 都是线性函数，不等式约束 $c_i, i = p+1, \ldots, p+q$ 都是凹函数，目标函数 $f(\boldsymbol{x})$ 是凸函数，则称(6.4)为**凸规划问题**。凸规划问题的可行域 D 是凸集，其局部极小值点一定是全局最小值点。

下面给出凸函数的另外一些性质。

(1) 若 T 为凸集，$\boldsymbol{y} \in \mathbb{R}^d$, 则 $f(\boldsymbol{x}) = \mathrm{dist}(\boldsymbol{y}, T)$ 是凸函数。

(2) 若 $f(\boldsymbol{x})$ 是 $\mathbb{R}^d$ 上的凸函数，$c \in \mathbb{R}$, 则 $\{\boldsymbol{x} \in \mathbb{R}^d : \ f(\boldsymbol{x}) \le c\}$ 和 $\{\boldsymbol{x} \in \mathbb{R}^d : \ f(\boldsymbol{x}) < c\}$ 都是凸集。

(3) $f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d$ 是凸函数当且仅当 $\{(\boldsymbol{x}, y) : \ f(\boldsymbol{x}) \le y, \boldsymbol{x} \in \mathbb{R}^d, y \in \mathbb{R}\}$ 是凸集。

(4) 若 $f(\boldsymbol{x})$ 是定义在 $\mathbb{R}^d$ 的开凸集 S 上的可微函数，则 $f(\boldsymbol{x})$ 是凸函数当且仅当

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \ge \nabla f(\boldsymbol{x})^T \ (\boldsymbol{y} - \boldsymbol{x}), \ \forall \boldsymbol{x}, \boldsymbol{y} \in S.$$

(5) 设 $f(x)$ 为定义在某区间 S 上的二阶可微的一元实函数，如果 $f(x)$ 满足 $f''(x) \ge 0, \forall x \in S$，则 $f(x)$ 在 S 上为凸函数；如果 $f(x)$ 满足 $f''(x) > 0, \forall x \in S$，则 $f(x)$ 在 S 上为严格凸函数。

(6) 设 $f(\boldsymbol{x})$ 是定义在开凸集 $S \subset \mathbb{R}^d$ 上的二阶可微函数，若 $\nabla^2 f(\boldsymbol{x}), \boldsymbol{x} \in S$ 都是非负定阵，则 $f(\boldsymbol{x})$ 为凸函数；若 $\nabla^2 f(\boldsymbol{x}), \boldsymbol{x} \in S$ 都是正定阵，则 $f(\boldsymbol{x})$ 为严格凸函数。

(7) 设 $f(\boldsymbol{x})$ 是定义在凸集 $S \subset \mathbb{R}^d$ 上的凸函数，则 $f(\boldsymbol{x})$ 在 S 的内核 $\mathring{S}$ 上连续且满足局部 Lipschitz 条件，即 $\forall \boldsymbol{x} \in \mathring{S}$, 存在常数 c 使得对 $\boldsymbol{x}$ 的一个邻域内的 $\boldsymbol{y}, \boldsymbol{z}$ 都有 $|f(\boldsymbol{y}) - f(\boldsymbol{z})| \le c\|\boldsymbol{y} - \boldsymbol{z}\|$。

(8) 对随机变量 X，若 $f(x)$ 是凸函数，EX 和 $Ef(X)$ 存在，则 $f(EX) \le Ef(X)$。这称为 Jenson 不等式。

设 $f(\boldsymbol{x})$ 是正值函数，如果 $\log f(\boldsymbol{x})$ 是凸函数，则称 $f(\boldsymbol{x})$ 为**对数凸函数**。凸函数以及对数凸函数之间的组合具有如下性质:

(1) 设 $f(\boldsymbol{x})$ 为凸函数，$g(t), t \in \mathbb{R}$ 为单调增的凸函数，则 $g(f(\boldsymbol{x}))$ 为凸函数。

(2) 设 A 为矩阵，$\boldsymbol{b}$ 为向量，$f(\boldsymbol{x})$ 为凸函数，则 $f(A\boldsymbol{x} + \boldsymbol{b})$ 为凸函数。

(3) 如果 $f(\boldsymbol{x}), g(\boldsymbol{x})$ 为凸函数，$\alpha \geq 0, \beta \geq 0$, 则 $\alpha f(\boldsymbol{x}) + \beta g(\boldsymbol{x})$ 为凸函数。

(4) 如果 $f(\boldsymbol{x}), g(\boldsymbol{x})$ 为凸函数，则 $\max(f(x), g(x))$ 为凸函数。

(5) 设 $f_n(\boldsymbol{x}), n = 1, 2, \ldots$ 都是凸函数且 $\lim_{n\to\infty} f_n(\boldsymbol{x})$ 存在，则 $f(\boldsymbol{x}) \triangleq \lim_{n\to\infty} f_n(\boldsymbol{x})$ 是凸函数。

(6) 对数凸函数一定也是凸函数。

(7) 设 $f(\boldsymbol{x})$ 为凸函数，$g(t), t \in \mathbb{R}$ 为单调增的对数凸函数，则 $g(f(\boldsymbol{x}))$ 是对数凸函数。

(8) 若 $f(\boldsymbol{x})$ 是对数凸函数，则 $f(A\boldsymbol{x} + \boldsymbol{b})$ 也是对数凸函数。

(9) 设 $f(\boldsymbol{x})$ 为对数凸函数，$\alpha > 0$，则 $f(\boldsymbol{x})^\alpha$ 和 $\alpha f(\boldsymbol{x})$ 为对数凸函数。

(10) 设 $f(\boldsymbol{x}), g(\boldsymbol{x})$ 为对数凸函数，则 $f(\boldsymbol{x}) + g(\boldsymbol{x})$, $f(\boldsymbol{x})g(\boldsymbol{x})$ 和 $\max(f(\boldsymbol{x}),\ g(\boldsymbol{x}))$ 都是对数凸函数。

(11) 设 $f_n(\boldsymbol{x}), n = 1, 2, \ldots$ 都是对数凸函数且 $\lim_{n\to\infty} f_n(\boldsymbol{x})$ 存在，则 $f(\boldsymbol{x}) \triangleq \lim_{n\to\infty} f_n(\boldsymbol{x})$ 是对数凸函数。

例 6.1.6. $f(x) = |x|^\gamma$ 是凸函数 $(\gamma \geq 1)$。首先, $g(x) = x^\gamma, x \in [0, \infty)$ 满足 $g''(x) = \gamma(\gamma - 1)x^{\gamma-2} \geq 0$，所以 $g(x)$ 在 $[0, \infty)$ 上是凸函数。于是，对任意的 $x, y \in (-\infty, \infty)$, $\alpha \in [0, 1]$ 有

$$\begin{aligned} f[\alpha x + (1-\alpha)y] =& |\alpha x + (1-\alpha)y|^\gamma \leq [\alpha|x| + (1-\alpha)|y|]^\gamma \\ =& g[\alpha|x| + (1-\alpha)|y|] \leq \alpha g(|x|) + (1-\alpha)g(|y|) \\ =& \alpha|x|^\gamma + (1-\alpha)|y|^\gamma = \alpha f(x) + (1-\alpha)f(y). \end{aligned}$$

□

例 6.1.7. 对线性函数 $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b, \boldsymbol{x} \in \mathbb{R}^n$，其中 $\boldsymbol{a} \in \mathbb{R}^d, b \in \mathbb{R}$, 有 $\nabla f(\boldsymbol{x}) = \boldsymbol{a}$, $\nabla^2 f(\boldsymbol{x}) = 0$，$f(\boldsymbol{x})$ 是凸函数，且(6.5)中的等式成立。 □

例 6.1.8. 对 $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x} + c, \boldsymbol{x} \in \mathbb{R}^n$, 其中 A 为非负定阵，$\boldsymbol{b} \in \mathbb{R}^d, c \in \mathbb{R}$, 有 $\nabla f(x) = A\boldsymbol{x} + \boldsymbol{b}$, $\nabla^2 f(\boldsymbol{x}) = A$, 所以 $f(\boldsymbol{x})$ 是凸函数。当 A 是正定阵时，$f(\boldsymbol{x})$ 是严格凸函数。 □

例 6.1.9. 欧式模 $f(\boldsymbol{x}) = \|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$ 是凸函数。事实上，当 $\boldsymbol{x} \neq \boldsymbol{0}$ 时 $\nabla f(\boldsymbol{x}) = \|\boldsymbol{x}\|^{-1}\boldsymbol{x}$, $\nabla^2 f(\boldsymbol{x}) = \|\boldsymbol{x}\|^{-3}\left(\|\boldsymbol{x}\|^2 I_d - \boldsymbol{x}\boldsymbol{x}^T\right)$, 其中 I_d 为 d 阶单位阵。易见 $\nabla^2 f(\boldsymbol{x})$ 是非负定阵，但不是正定阵。

下面直接验证 $f(\boldsymbol{x})=\|\boldsymbol{x}\|$ 是凸函数。对任意 $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d$, $\alpha\in(0,1)$，有

$$\begin{aligned} f[\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}] =&\|\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}\|\le\|\alpha\boldsymbol{x}\|+\|(1-\alpha)\boldsymbol{y}\| \\ =&\alpha\|\boldsymbol{x}\|+(1-\alpha)\|\boldsymbol{y}\|=\alpha f(\boldsymbol{x})+(1-\alpha)f(\boldsymbol{y}) \end{aligned}$$

成立。 □

例 6.1.10. 设 $\boldsymbol{z}_j, j=1,\dots,n$ 是 $\mathbb{R}^d$ 的 n 个点。定义

$$f(\boldsymbol{x})=\log\left[\sum_{j=1}^{n}\exp(\boldsymbol{z}_j^T\boldsymbol{x})\right],$$

由于 $\boldsymbol{z}_j^T\boldsymbol{x}$ 是凸函数，e^t 是单调增对数凸函数，所以 $\exp(\boldsymbol{z}_j^T\boldsymbol{x})$ 是对数凸函数，从而 $f(\boldsymbol{x})$ 是凸函数。 □

6.1.4 无约束极值点的条件

设 $f(\boldsymbol{x})$ 的定义域 $\mathcal{A}$ 为 $\mathbb{R}^d$ 中的开集。如果 $\boldsymbol{x}^*$ 使得 $\nabla f(\boldsymbol{x}^*)=0$，称 $\boldsymbol{x}^*$ 为 $f(\cdot)$ 的一个**稳定点**。无约束极值点的必要条件和充分条件是数学分析中熟知的结论，这里仅罗列这些结论备查。

定理 6.1.1 (一阶必要条件). 如果 $f(\boldsymbol{x}),\boldsymbol{x}\in\mathcal{A}$ 有一阶连续偏导数，$\boldsymbol{x}^*\in\mathcal{A}$ 是 $f(\cdot)$ 的一个局部极小值点，则 $\boldsymbol{x}^*$ 是 $f(\cdot)$ 的稳定点：

$$\nabla f(\boldsymbol{x}^*)=\boldsymbol{0}. \tag{6.6}$$

定理 6.1.2 (二阶必要条件). 如果 $f(\boldsymbol{x}),\boldsymbol{x}\in\mathcal{A}$ 有二阶连续偏导数，$\boldsymbol{x}^*\in\mathcal{A}$ 是 $f(\cdot)$ 的一个局部极小值点，则 $\boldsymbol{x}^*$ 是 $f(\cdot)$ 的稳定点，且海色阵 $\nabla^2 f(\boldsymbol{x}^*)$ 为非负定阵。

定理 6.1.3 (二阶充分条件). 如果 $f(\boldsymbol{x}),\boldsymbol{x}\in\mathcal{A}$ 有二阶连续偏导数，$\boldsymbol{x}^*\in\mathcal{A}$ 是 $f(\cdot)$ 的一个稳定点，且海色阵 $\nabla^2 f(\boldsymbol{x}^*)$ 为正定阵，则 $\boldsymbol{x}^*$ 是 $f(\cdot)$ 的局部严格极小值点。

稳定点不一定是极值点。如果 $\boldsymbol{x}^*$ 是稳定点但是 $\nabla^2 f(\boldsymbol{x}^*)$ 同时有正特征值和负特征值，则 $\boldsymbol{x}^*$ 一定不是 $f(\cdot)$ 的极值点；如果 $\boldsymbol{x}^*$ 是稳定点而 $\nabla^2 f(\boldsymbol{x}^*)$ 非负定但不正定，这时 $\boldsymbol{x}^*$ 可能是 $f(\cdot)$ 的极值点，也可能不是。

当所有 $\nabla^2 f(\boldsymbol{x})$ 都是非负定（也称半正定）矩阵时，$f(\boldsymbol{x})$ 是凸函数；当所有 $\nabla^2 f(\boldsymbol{x})$ 都是正定矩阵时，$f(\boldsymbol{x})$ 是严格凸函数。由凸函数的性质可得如下结论。

定理 6.1.4 (全局最小值点的二阶充分条件). 如果 $f(\boldsymbol{x}),\boldsymbol{x}\in\mathcal{A}$ 有二阶连续偏导数，所有 $\nabla^2 f(\boldsymbol{x})$ 都是非负定阵，$\boldsymbol{x}^*\in\mathcal{A}$ 是 $f(\cdot)$ 的一个稳定点，则 $\boldsymbol{x}^*$ 是 $f(\cdot)$ 的全局最小值点; 如果进一步设 $\nabla^2 f(\boldsymbol{x}^*)$ 是正定阵，则稳定点 $\boldsymbol{x}^*$ 是全局严格最小值点。

定义 (方向导数) 对非零向量 $\boldsymbol{v} \in \mathbb{R}^d$, 定义一元函数

$$h(\alpha) = f\left(\boldsymbol{x} + \alpha \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}\right),\ \alpha \geq 0 \tag{6.7}$$

如果 $h(\alpha)$ 在 $\alpha = 0$ 的右导数存在，则称其为函数 $f(\boldsymbol{x})$ 在点 $\boldsymbol{x}$ 延方向 $\boldsymbol{v}$ 的**方向导数**，记为 $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{v}}$；类似可定义二阶方向导数 $\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{v}^2}$。

如果 $f(\boldsymbol{x})$ 有一阶连续偏导数，$\|\boldsymbol{v}\| = 1$, 则 $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{v}} = \boldsymbol{v}^T \nabla f(\boldsymbol{x})$。如果 $f(\boldsymbol{x})$ 有二阶连续偏导数，$\|\boldsymbol{v}\| = 1$, 则 $\frac{\partial^2 f(\boldsymbol{x})}{\partial \boldsymbol{v}^2} = \boldsymbol{v}^T \nabla^2 f(\boldsymbol{x}) \boldsymbol{v}$。

求无约束极值通常使用迭代法。设迭代中得到了一个近似极小值点 $\boldsymbol{x}^{(t)}$，如果 $\nabla f(\boldsymbol{x}) \neq \boldsymbol{0}$，则从 $\boldsymbol{x}^{(t)}$ 出发延 $-\nabla f(\boldsymbol{x})$ 方向前进可以使函数值下降。而且，对任意非零向量 $\boldsymbol{v} \in \mathbb{R}^d$, 只要 $\boldsymbol{v}^T \nabla f(\boldsymbol{x}) < 0$，则 $\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{v}} < 0$, 从 $\boldsymbol{x}^{(t)}$ 出发延 $\boldsymbol{v}$ 方向前进也可以使函数值下降，称这样的方向 $\boldsymbol{v}$ 为**下降方向**。

例 6.1.11. 设 $\boldsymbol{Y}$ 为 $n \times 1$ 向量, $\boldsymbol{X}$ 为 $n \times p$ 矩阵 $(n > p)$，$\boldsymbol{\beta}$ 为 $p \times 1$ 向量。求函数

$$Q(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{6.8}$$

的极小值点的问题称为线性最小二乘问题。易见

$$\nabla Q(\boldsymbol{\beta}) = -2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, \tag{6.9}$$

$$\nabla^2 Q(\boldsymbol{\beta}) = 2\boldsymbol{X}^T\boldsymbol{X}, \tag{6.10}$$

当 $\boldsymbol{X}$ 列满秩时 $\boldsymbol{X}^T\boldsymbol{X}$ 为正定矩阵，这时稳定点 $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y}$ 是 $Q(\boldsymbol{\beta})$ 全局严格最小值点。

如果 X 不满秩，因为 X^TX 是非负定矩阵，所以 $Q(\boldsymbol{\beta})$ 是凸函数，可以证明 $Q(\boldsymbol{\beta})$ 存在无穷多个稳定点，由定理6.1.4可知这些稳定点都是全局最小值点 (见习题5)。 □

例 6.1.12. 考虑

$$f(\boldsymbol{x}) = \frac{1}{4}x_1^4 + \frac{1}{2}x_2^2 - x_1x_2 + x_1 - x_2,\ \boldsymbol{x} \in \mathbb{R}^2,$$

易见

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} x_1^3 - x_2 + 1 \\ x_2 - x_1 - 1 \end{pmatrix}, \quad \nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} 3x_1^2 & -1 \\ -1 & 1 \end{pmatrix},$$

令 $\nabla f(\boldsymbol{x}) = 0$ 得到三个稳定点 $(-1,0), (1,2), (0,1)$。计算这三个点处的海色阵得

$$\nabla^2 f(-1,0) = \nabla^2 f(1,2) = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}, \quad \nabla^2 f(0,1) = \begin{pmatrix} 0 & -1 \\ -1 & 1 \end{pmatrix},$$

前两个稳定点海色阵的特征值为 $2\pm\sqrt{2}>0$，是正定阵，所以 $(-1,0)$ 和 $(1,2)$ 是 $f(\cdot)$ 的极小值点，$f(-1,0)=f(1,2)=-\frac{3}{4}$，这两个点都是全局最小值点；稳定点 $(0,1)$ 处的海色阵的特征值为 $\frac{1}{2}(1\pm\sqrt{5})$，一正一负，所以 $(0,1)$ 不是极小值点，且 $f(0,1)=-\frac{1}{2}$。 □

6.1.5 约束极值点的条件 *

与无约束最优化问题不同，约束极值点一般不满足梯度向量为零的条件，这是因为约束极值经常在可行域的边界达到，这时在极值点处如果不考虑约束条件，函数值仍可下降，所以约束极值点处的梯度不需要等于零向量。

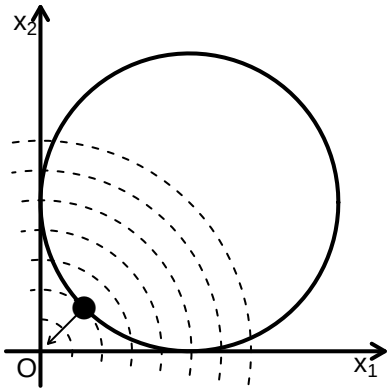

图 6.1: 例6.1.13图形。虚线为 $f(\boldsymbol{x})$ 的等值线，箭头是约束最小值点的负梯度方向

例 6.1.13. 考虑如下约束优化问题

$$\begin{cases}\arg\min f(\boldsymbol{x})=x_1^2+x_2^2,\ \text{s.t.}\\ (x_1-1)^2+(x_2-1)^2\le 1,\end{cases}\tag{6.11}$$

可行域 D 是以 $(1,1)^T$ 为圆心的半径等于 1 的圆面。容易看出约束最小值在 $f(\boldsymbol{x})$ 的等值线与 D 外切的点 $\boldsymbol{x}^*=\left(1-\frac{\sqrt{2}}{2},1-\frac{\sqrt{2}}{2}\right)^T$ 处达到 (见图6.1)，这时 $\nabla f(\boldsymbol{x}^*)=(2-\sqrt{2},2-\sqrt{2})^T\neq\mathbf{0}$，沿着其反方向 $(-1,-1)^T$ 搜索仍可使函数值下降，但会离开可行域。 □

在仅有等式约束的情形，如下的拉格朗日乘子法给出了约束极值点的必要条件。

定理 6.1.5. 考虑约束最优化问题

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}), \text{ s.t.} \\ c_i(\boldsymbol{x}) = 0,\ i = 1,\ldots,p \end{cases} \tag{6.12}$$

设 $f(\cdot), c_i(\cdot)$ 在 $\mathbb{R}^d$ 存在一阶连续偏导数，定义拉格朗日函数

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{i=1}^{p}\lambda_i c_i(\boldsymbol{x}), \tag{6.13}$$

其中 $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_p)^T$，若 $\boldsymbol{x}^*$ 是(6.12)的一个约束局部极小值点，则必存在 $\boldsymbol{\lambda}^*$ 使得 $(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 为 $\mathcal{L}(\boldsymbol{x},\boldsymbol{\lambda})$ 的稳定点, 即

$$\nabla f(\boldsymbol{x}^*) - \sum_{i=1}^{p}\lambda_i^*\nabla c_i(\boldsymbol{x}^*) = 0, \tag{6.14}$$

$$c_i(\boldsymbol{x}^*) = 0,\ i = 1,2,\ldots,p. \tag{6.15}$$

条件(6.14)说明在约束局部极小值点，目标函数的梯度是各个约束的梯度的线性组合。

考虑如(6.4)那样的一般约束最优化问题。设可行域 D 为非空集。关于不等式约束 $c_i, i = p+1,\ldots,p+q$，对 $\boldsymbol{x}\in D$，如果 $c_i(\boldsymbol{x}) = 0$，称 c_i 在 $\boldsymbol{x}$ 处是**起作用**的，否则称 c_i 在 $\boldsymbol{x}$ 处是**不起作用**的。实际上，假设 f, c_i 都连续，如果 $\boldsymbol{x}^*$ 是一个约束局部极小值点，则去掉 $\boldsymbol{x}^*$ 处不起作用的约束得到的新优化问题也以 $\boldsymbol{x}^*$ 为一个约束局部极小值点。起作用的和不起作用的不等式约束的下标集合是随 $\boldsymbol{x}$ 而变的，在本书中，为记号简单起见，设 $c_i, i = p+1,\ldots,p+r(0\le r\le q)$ 是起作用的不等式约束，$c_i, i = p+r+1,\ldots,p+q$ 是不起作用的不等式约束，对实际问题则应使用两个随 $\boldsymbol{x}$ 而变化的下标集合或每次重新排列约束次序。

由于可行域形状可能是多种多样的，需要特别地定义可行方向的概念。

定义（可行方向）设 $\boldsymbol{x}$ 是问题(6.4)的可行点，如果有 $\boldsymbol{d}\in\mathbb{R}^d, \|\boldsymbol{d}\| = 1$, 以及序列 $\boldsymbol{d}^{(k)}\in\mathbb{R}^d$, $\|\boldsymbol{d}^{(k)}\| = 1$, 和 $\alpha_k > 0$，使得 $\boldsymbol{x}^{(k)} = \boldsymbol{x} + \alpha_k\boldsymbol{d}^{(k)}\in D$, 且 $\lim_{k\to\infty}\boldsymbol{x}^{(k)} = \boldsymbol{x}$，$\lim_{k\to\infty}\boldsymbol{d}^{(k)}\to\boldsymbol{d}$, 则称 $\boldsymbol{d}$ 是问题(6.4)在 $\boldsymbol{x}$ 处的一个**可行方向**，记 $\boldsymbol{x}$ 处所有可行方向的集合为 $\mathscr{F}(\boldsymbol{x})$。可行方向也称为**序列可行方向**。如果 $\boldsymbol{d}\in\mathscr{F}(\boldsymbol{x})$ 满足 $\boldsymbol{d}^T\nabla f(\boldsymbol{x}) < 0$，称 $\boldsymbol{d}$ 是 $\boldsymbol{x}$ 处的一个**可行下降方向**。

另一种可行方向定义比较简单。

定义（线性化可行方向） 设 $\boldsymbol{x}$ 是问题(6.4)的可行点，

$$\begin{aligned}\mathscr{L}(\boldsymbol{x}) \triangleq \big\{\boldsymbol{d}\in\mathbb{R}^d : \|\boldsymbol{d}\| = 1,\ \boldsymbol{d}^T\nabla c_i(\boldsymbol{x}) = 0, i = 1,\ldots,p; \\ \boldsymbol{d}^T\nabla c_i(\boldsymbol{x}) \ge 0, i = p+1,\ldots,p+r\big\}\end{aligned} \tag{6.16}$$

称为 $\boldsymbol{x}$ 处的线性化可行方向集，其中的元素称为 $\boldsymbol{x}$ 处的**线性化可行方向**。

注意线性化可行方向定义只对等式约束和起作用的不等式约束有要求，不关心不起作用的不等式约束。可以证明，序列可行方向一定是线性化可行方向，反之不然。在线性化可行方向定义中，要求 $\boldsymbol{d}$ 与等式约束 $c_i(\boldsymbol{x})$ 的梯度正交，所以延 $\boldsymbol{d}$ 略微移动使得 $c_i(\boldsymbol{x})$ 既不增加也不减少，保持等式不变；要求 $\boldsymbol{d}$ 与起作用的不等式约束 $c_i(\boldsymbol{x})$ 的梯度不能反向，所以延 $\boldsymbol{d}$ 略微移动不能使得 $c_i(\boldsymbol{x})$ 减小，所以可以保持不等式成立。

如果 $\boldsymbol{x}^*$ 是问题(6.4)的一个约束局部极小值点，则 $\boldsymbol{x}^*$ 处没有可行下降方向，否则在 $\boldsymbol{x}^*$ 附近可以找到函数值比 $f(\boldsymbol{x}^*)$ 更小的可行点。另一方面，如果可行点 $\boldsymbol{x}^*$ 处所有可行方向 $\boldsymbol{d} \in \mathscr{F}(\boldsymbol{x}^*)$ 都是上升方向 (即 $\boldsymbol{d}^T\nabla f(\boldsymbol{x}) > 0$)，则 $\boldsymbol{x}^*$ 是一个约束严格局部极小值点。

鉴于可行方向集难以确定，以下的定理给出了约束极值点的一个比较容易验证的必要条件。

定理 6.1.6 (Karush-Kuhn-Tucker). 对约束最优化问题(6.4)，假定目标函数 $f(\boldsymbol{x})$ 和约束函数 $c_i(\boldsymbol{x})$ 在 $\mathbb{R}^d$ 有一阶连续偏导数，$\boldsymbol{x}^*$ 是一个约束局部极小值点，如果 $c_i, i = 1, \ldots, p+r$ 都是线性函数，或者 $\{\nabla c_i(\boldsymbol{x}^*), i = 1, \ldots, p+r\}$ 构成线性无关向量组，则必存在常数 $\lambda_1^*, \ldots, \lambda_{p+q}^*$，使得

$$\nabla f(\boldsymbol{x}^*) - \sum_{i=1}^{p+q} \lambda_i^* \nabla c_i(\boldsymbol{x}^*) = 0, \tag{6.17}$$

$$\lambda_i^* \geq 0, \ i = p+1, \ldots, p+q, \tag{6.18}$$

$$\lambda_i^* c_i(\boldsymbol{x}^*) = 0, \ i = p+1, \ldots, p+q. \tag{6.19}$$

对问题(6.4)仍可定义拉格朗日函数

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) - \sum_{i=1}^{p+q} \lambda_i c_i(\boldsymbol{x}), \tag{6.20}$$

定理结论中的(6.19)表明 $L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = f(\boldsymbol{x}^*)$, (6.17)可以写成 $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = 0$, 其中 $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{\lambda})$ 表示 $L(\boldsymbol{x}, \boldsymbol{\lambda})$ 的梯度向量中与分量 $\boldsymbol{x}$ 对应的部分。

定理证明略去 (见高立 (2014)[3]§6.3, Lange(2013)[25]§5.2)。(6.17)–(6.19)称为 KKT **条件**或 KT 条件，迭代时满足 KKT 条件的近似最小值点 $\boldsymbol{x}^*$ 称为 KKT **点**, 对应的 $\boldsymbol{\lambda}^*$ 称为**拉格朗日乘子**，$(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ 称为 KKT **对**。KKT 点类似于无约束优化问题中的稳定点，很多基于导数的算法在达到 KKT 点后就不能再继续搜索。

注意(6.19)保证了不起作用的不等式约束对应的乘子 $\lambda_i^* = 0, i = p+r+1, \ldots, p+q$。这样，我们把 KKT 对 $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ 处的约束条件 $c_i(\boldsymbol{x}^*), i = 1, \ldots, p+q$ 和相应的拉格朗日乘子 $\boldsymbol{\lambda}^*$ 分为四个部分:

(1) $c_i(\boldsymbol{x}^*)=0$, λ_i^* 可取正、负、零, $i=1,\ldots,p$;

(2) $c_i(\boldsymbol{x}^*)=0$, $\lambda_i^*>0$, $i=1,\ldots,p+r'(r'\le r)$;

(3) $c_i(\boldsymbol{x}^*)=0$, $\lambda_i^*=0$, $i=p+r'+1,\ldots,p+r$;

(4) $c_i(\boldsymbol{x}^*)>0$, $\lambda_i^*=0$, $i=p+r+1,\ldots,p+q$。

其中，第一部分对应于等式约束，第二、第三部分对应于起作用的不等式约束，第四部分对应于不起作用的不等式约束。注意，这里为了记号简单起见用序号分开了这四部分，实际上第二、三、四部分的组成下标可以是 $\{p+1,\ldots,p+q\}$ 的任意分组。把(6.17)写成

$$\nabla f(\boldsymbol{x}^*)=\sum_{i=1}^{p+r'}\lambda_i^*\nabla c_i(\boldsymbol{x}^*), \tag{6.21}$$

对线性化可行方向 $\boldsymbol{d}\in\mathscr{L}(\boldsymbol{x}^*)$, 有

$$\boldsymbol{d}^T\nabla f(\boldsymbol{x}^*)=\sum_{i=1}^{p}\lambda_i^*\boldsymbol{d}^T\nabla c_i(\boldsymbol{x}^*)+\sum_{i=p+1}^{p+r'}\lambda_i^*\boldsymbol{d}^T\nabla c_i(\boldsymbol{x}^*)\ge 0, \tag{6.22}$$

所以在 KKT 点没有可行下降方向; 没有可行下降方向是约束极小值点的必要条件，但不是充分条件, 对于那些与 $\nabla f(\boldsymbol{x}^*)$ 正交的可行方向，需要进一步的信息来判断。

在 KKT 对 $(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 处定义

$$\begin{aligned}\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)\triangleq\big\{\boldsymbol{d}\in\mathbb{R}^d:\|\boldsymbol{d}\|=1,\boldsymbol{d}^T\nabla c_i(\boldsymbol{x}^*)=0,\ i=1,\ldots,p+r';\\ \boldsymbol{d}^T\nabla c_i(\boldsymbol{x}^*)\ge 0,i=p+r'+1,\ldots,p+r\big\},\end{aligned} \tag{6.23}$$

显然 $\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)\subset\mathscr{L}(\boldsymbol{x}^*)$, 且对 $\boldsymbol{d}\in\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 有

$$\boldsymbol{d}^T\nabla f(\boldsymbol{x}^*)=\sum_{i=1}^{p+r'}\lambda_i^*\boldsymbol{d}^T\nabla c_i(\boldsymbol{x}^*)=0, \tag{6.24}$$

即 $\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 是 $\mathscr{L}(\boldsymbol{x}^*)$ 中与梯度 $\nabla f(\boldsymbol{x}^*)$ 正交的方向。因为这样的方向可能存在，判断极小值点可能还需要二阶偏导数条件。这个问题的理由与一元函数导数等于零的点不一定是极值点的理由类似。

定理 6.1.7 (二阶必要条件). 在定理6.1.6的条件下，进一步设在 $\boldsymbol{x}^*$ 的一个邻域内 f 与各 c_i 有二阶连续偏导数，则有

$$\boldsymbol{d}^T\nabla^2_{\boldsymbol{xx}}L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)\boldsymbol{d}\ge 0,\ \forall\boldsymbol{d}\in\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*). \tag{6.25}$$

定理6.1.7给出了约束极小值点处拉格朗日函数关于 $\boldsymbol{x}$ 的海色阵的必要条件，类似于无约束情形下要求目标函数海色阵非负定，(6.25)针对方向 $\boldsymbol{d}\in\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 要求 $\nabla^2_{\boldsymbol{xx}}L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 非负定。$\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 是 $\mathscr{L}(\boldsymbol{x}^*)$ 中与梯度 $\nabla f(\boldsymbol{x}^*)$ 正交的方向，为了 $\boldsymbol{x}^*$ 是局部最小值点，对这样的方向需要加二阶导数条件，因为在这样的方向上变化仅从一阶导数看不出会不会下降。

定理 6.1.8 (二阶充分条件). 对约束最优化问题(6.4)，假定目标函数 f 和各约束函数 c_i 在 $\boldsymbol{x}^*$ 的邻域内有二阶连续偏导数，若 $(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 是 KKT 对，且

$$\boldsymbol{d}^T\nabla^2_{\boldsymbol{xx}}L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)\boldsymbol{d}>0,\ \forall\boldsymbol{d}\in\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*), \tag{6.26}$$

则 $\boldsymbol{x}^*$ 是问题(6.4)的一个严格局部极小值点。

如果在 KKT 对 $(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 处的 $\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 为空集, 定理6.1.8结论成立。

定理6.1.7和定理6.1.8的证明参见高立 (2014)[3] §6.4。

例 6.1.14. 利用 Karush-Kuhn-Tucker 定理可以证明如下不等式：

$$\frac{1}{4}(x_1^2+x_2^2)\le e^{x_1+x_2-2},\ x_1\ge 0,x_2\ge 0. \tag{6.27}$$

只要证明 $f(\boldsymbol{x})=-(x_1^2+x_2^2)e^{-x_1-x_2}$ 的最小值是 $-4e^{-2}$。取 $p=0,q=2$, 约束函数 $c_1(\boldsymbol{x})=x_1$, $c_2(\boldsymbol{x})=x_2$, 约束函数都是线性函数，极小值点应该满足条件

$$\begin{aligned}
&e^{-x_1-x_2}\begin{pmatrix}-2x_1+x_1^2+x_2^2\\ -2x_2+x_1^2+x_2^2\end{pmatrix}-\lambda_1\begin{pmatrix}1\\0\end{pmatrix}-\lambda_2\begin{pmatrix}0\\1\end{pmatrix}=0,\\
&\lambda_1\ge 0,\ \lambda_2\ge 0,\\
&\lambda_1x_1=0,\ \lambda_1x_2=0.
\end{aligned}$$

解集为 $\{(0,0),(1,1),(0,2),(2,0)\}$, 对应的目标函数值分别为 0, $-2e^{-2}$, $-4e^{-2}$, $-4e^{-2}$。易见 $\lim_{\|\boldsymbol{x}\|\to\infty}f(\boldsymbol{x})=0>-4e^{-2}$, 所以目标函数 $f(\boldsymbol{x})$ 能达到最小值, 所以全局约束最小值点是 $(0,2)$ 和 $(2,0)$，最小值是 $-4e^{-2}$。

作为示例，用二阶条件来判断 $\boldsymbol{x}^*=(2,0)^T$ 是否局部极小值点。解出对应的拉格朗日乘子为 $\lambda^*=(0,4e^{-2})^T$, 两个约束条件中第二个起作用且对应的拉格朗日乘子为正，第一个不起作用，又 $\nabla c_1(\boldsymbol{x}^*)=(1,0)^T$, $\nabla c_2(\boldsymbol{x}^*)=(0,1)^T$,

$$\mathscr{L}_1(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)=\{\boldsymbol{d}:\ \|\boldsymbol{d}\|=1,\boldsymbol{d}^T(1,0)^T\ge 0,\boldsymbol{d}^T(0,1)^T=0\}=\{(1,0)^T\},$$

计算可得 $(1,0)\nabla^2L_{\boldsymbol{xx}}(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)(1,0)^T=2>0$，所以 $\boldsymbol{x}^*=(2,0)^T$ 是局部极小值点。 □

6.1.6 迭代收敛

最优化和方程求根普遍使用迭代算法，从上一步的近似值 $\boldsymbol{x}^{(t)}$ 通过迭代公式得到下一步的近似值 $\boldsymbol{x}^{(t+1)}$。设 $\lim_{t\to\infty}\boldsymbol{x}^{(t)}=\boldsymbol{x}^*$，下面给出收敛速度的一些度量。

定义 令

$$Q_1=\varlimsup_{t\to\infty}\frac{\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\|}, \tag{6.28}$$

则 $Q_1=0$ 时称算法**超线性收敛**, 当 $0<Q_1<1$ 时称算法**线性收敛**, 当 $Q_1=1$ 时称算法**次线性收敛**。

定义 令

$$Q_2=\varlimsup_{t\to\infty}\frac{\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^*\|}{\|\boldsymbol{x}^{(t)}-\boldsymbol{x}^*\|^2}, \tag{6.29}$$

则 $Q_2=0$ 时称算法**超平方收敛**, 当 $0<Q_2<+\infty$ 时称算法**平方收敛**或**二次收敛**, 当 $Q_2=+\infty$ 时称算法**次平方收敛**。

最优化算法至少应该有线性收敛速度，否则收敛太慢，实际中无法使用。

实际的最优化问题不一定满足算法收敛的条件，而且往往难以预先判断是否能够收敛。所以，最优化算法对于何时停止迭代，通常使用目标函数值 $f(\boldsymbol{x}^{(t)})$ 两次迭代之间的变化量 $|f(\boldsymbol{x}^{(t+1)})-f(\boldsymbol{x}^{(t)})|$ 小于预定精度值、迭代近似点 $\boldsymbol{x}^{(t)}$ 两次之间的变化量 $\|\boldsymbol{x}^{(t+1)}-\boldsymbol{x}^{(t)}\|$ 小于预定精度值、梯度向量长度 $\|\nabla f(\boldsymbol{x}^{(t)})\|$ 小于预定精度值等作为停止法则，另外，为了保险起见当迭代次数超出一个预定最大次数时也停止，但是这时给出算法失败的结果。

在统计计算问题中，目标函数 $f(\boldsymbol{x})$ 常常不能计算到很高精度，有时目标函数值甚至是由随机模拟方法计算得到的，这样，迭代停止法则不能选取过高的精度，否则算法可能在极值点附近不停跳跃而无法收敛。

6.2 一维搜索与求根

即使目标函数 $f(x)$ 是一元函数，求最小值点也经常需要使用数值迭代方法。另外，在多元目标函数优化中，一般每次迭代从上一步的 $\boldsymbol{x}^{(t)}$ 先确定一个下降方向 $\boldsymbol{d}^{(t)}$, 然后对派生出的一元函数 $h(\alpha)=f(\boldsymbol{x}^{(t)}+\alpha\boldsymbol{d}^{(t)})$, $\alpha\geq 0$ 求最小值点得到下降的步长 α_t，并令 $\boldsymbol{x}^{(t+1)}=\boldsymbol{x}^{(t)}+\alpha_t\boldsymbol{d}^{(t)}$，求步长的过程称为**一维搜索**，搜索可以是求一元函数 $h(\alpha)$ 的精确最小值点，也可以求一个使得目标函数下降足够多的 α 作为步长。对于多元目标函数 $f(\boldsymbol{x})$, $\boldsymbol{x}\in\mathbb{R}^d$, 有一种优化方法是先延 x_1 坐标轴方向搜索，再延 x_2 坐标轴方向搜索，直到延 x_d 坐标轴方向搜索后再重复地从 x_1 坐标轴方向开始，这种方法叫做**坐标循环下降法**。

下面讨论一元函数的优化和求根问题。

6.2.1 二分法求根

优化问题与求根问题密切相关，很多优化问题会归结为一个求根问题。对一元目标函数 $f(x)$，若 $f(x)$ 连续可微，极值点一定是 $f'(x)=0$ 的根。二分法是实际数值计算中一元函数求根使用最多的方法，这种方法简单而又有比较高的收敛速度。

设函数 $f(x)$ 在闭区间 $[a,b]$ 连续，且 $f(a)f(b)\le 0$，由中间值定理，至少有一个点 $x^*\in[a,b]$ 使得 $f(x^*)=0$。如果 $f(x)$ 在 $[a,b]$ 上还是严格单调函数则解 x^* 存在唯一。

二分法求根用区间 $[a,b]$ 中点 c 处函数值 $f(c)$ 的正负号来判断根在区间中点的左边还是右边，显然，如果 $f(c)$ 与 $f(a)$ 异号，则 $[a,c]$ 中有根；如果 $f(c)$ 与 $f(b)$ 异号，则 $[c,b]$ 中有根。根据这样的想法，可以迭代地每次用区间的左半部分或右半部分代替原来的含根区间，逐步缩小含根的区间。设求根的绝对误差限规定为 ϵ。算法如下:

令 $f_a=f(a),\ f_b=f(b)$
until $(b-a<\epsilon)$ {
    令 $c\leftarrow\frac{1}{2}(a+b)$, $f_c\leftarrow f(c)$
    if $(f_af_c\le 0)$ {
        $b\leftarrow c$, $f_b\leftarrow f_c$
    } **else** {
        $a\leftarrow c$, $f_a\leftarrow f_c$
    }
}
输出 c 作为根

设真实根为 x^*，第 k 步的区间中点为 $x^{(k)}$，则

$$|x^{(k)}-x^*|\le(b-a)\left(\frac{1}{2}\right)^k,$$

二分法具有线性收敛速度。

如果 $f(x)$ 是 $[a,\infty)$ 区间上的严格单调的连续函数, 且 $\lim_{x\to\infty}f(x)$ 与 $f(a)$ 反号，则 $f(x)$ 在 $[a,\infty)$ 上存在唯一的根，这时需要先找到闭区间 $[a,b]$ 使得 $f(a)$ 与 $f(b)$ 反号，二分法算法需要略作调整:

取步长 $h>0$, 倍数 $\gamma>1$
until $(f(a)f(b)\le 0)$ {
    $b\leftarrow a+h$
    $h\leftarrow\gamma h$

}

用闭区间上的二分法在 $[a,b]$ 内求根

其它的情况，比如 $f(x)$ 定义在 $(-\infty,\infty)$, (a,b) 的情况可类似处理，都需要先找到使得区间端点函数值符号相反的闭区间。

例 6.2.1. 设某种建筑钢梁的强度 X 服从正态 $\mathrm{N}(\mu,\sigma^2)$ 分布，μ,σ^2 未知，根据设计要求，当 $X>L(L$ 已知) 时钢梁的强度才达到要求，称 L 为强度的下规范限，称强度达到要求的概率 $R=P(X>L)$ 为单侧性能可靠度, 在可靠性评估中需要在给了总体 X 的一组样本 $X_1,X_2,\ldots,X_n$ 以后求可靠度 R 的 $1-\alpha$ 置信下限 $(0<\alpha<1)$。

记 $K=(\mu-L)/\sigma$，则 $R=\Phi(K)$, 所以 R 是 K 的严格单调增连续函数，只要求 K 的置信下限。

假设从样本 $X_1,X_2,\ldots,X_n$ 中计算了样本平均值 $\bar{X}$ 和样本方差 S^2，定义 $\hat{K}\triangleq(\bar{X}-L)/S$，$\hat{K}$ 的分布仅依赖于 K，经过简单推导可以得知 $\sqrt{n}\hat{K}$ 服从非中心 $\mathrm{t}(n-1,\sqrt{n}K)$ 分布, 设其分布函数为 $\mathrm{pt}(\cdot,n-1,\sqrt{n}K)$。

按照置信下限的统计量法 (见陈家鼎等 (2006)[1] §2.3)，只要定义

$$\begin{aligned}G(u,K)=&P(\hat{K}\geq u),\ u\in(0,1),\ K\in\mathbb{R},\\ g(u)=&\inf_{K\in(-\infty,\infty)}\{K:\ G(u,K)>\alpha\},\ u\in(0,1),\end{aligned}$$

则 $g(\hat{K})$ 是 K 的 $1-\alpha$ 置信下限, 从而 $\Phi(g(\hat{K}))$ 是 R 的置信下限。而

$$G(u,K)=1-\mathrm{pt}(\sqrt{n}u,n-1,\sqrt{n}K)$$

是 K 的严格单调增函数，且 $\lim_{K\to-\infty}G(u,K)=0$, $\lim_{K\to+\infty}G(u,K)=1$, 所以在 $G(u,K)>\alpha$ 约束下求 K 的下确界，等价于在固定 u 的情况下对关于 K 严格单调增的连续函数 $f(K)\triangleq G(u,K)-\alpha$ 求根，易见根存在唯一，只要用二分法求根即可。需要先求得包含根的闭区间。

求包含根的区间 $[a,b]$ 算法如下:

取初始步长 $h_0>0$, 倍数 $\gamma>1$

$h\leftarrow h_0$, $a\leftarrow 0$, $b\leftarrow 0$

while $(f(a)>0)$ { # 这时不需要单独找 b

$b\leftarrow a$

$a\leftarrow a-h$

$h\leftarrow\gamma h$

}
$h \leftarrow h_0$
while $(f(b) \le 0)$ { # 这时不需要单独找 a
$a \leftarrow b$
$b \leftarrow b + h$
$h \leftarrow \gamma h$
}
用闭区间上的二分法在 $[a, b]$ 内求根

□

6.2.2 牛顿法

设目标函数 $f(x)$ 有连续二阶导数，求最小值点可以通过求解 $f'(x) = 0$ 的根实现。记 $g(x) = f'(x)$。设 x^* 是 $f(x)$ 的一个最小值点，则 $g(x^*) = 0$，设 x_0 是 x^* 附近的一个点。在 x_0 附近对 $g(x)$ 作一阶泰勒近似得

$$g(x) \approx g(x_0) + g'(x_0)(x - x_0),$$

令 $g(x) = 0$，得 x 近似为

$$x_1 = x_0 - \frac{g(x_0)}{g'(x_0)},$$

写成迭代公式，即

$$x_{t+1} = x_t - \frac{g(x_t)}{g'(x_t)},\ t = 0, 1, 2, \ldots \tag{6.30}$$

或

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)},\ t = 0, 1, 2, \ldots \tag{6.31}$$

如果 $g'(x)$ 在 x^* 的一个邻域 $[x^* - \delta, x^* + \delta]$ 内都为正，初值 $x_0 \in [x^* - \delta, x^* + \delta]$, 则牛顿法产生的迭代数列 $\{x_t\}$ 收敛到 x^*。如果初值接近于最小值点并且目标函数满足一定正则性条件，牛顿法有二阶收敛速度。但是，如果迭代过程中遇到 $g'(x)$ 接近于零的点，下一个迭代点可能会被抛到远离根 x^* 的地方，造成不收敛。另外，牛顿法迭代的停止法则，可以取为 $|x_{t+1} - x_t| < \epsilon_x$ 或 $|g(x_t)| < \epsilon_g$, ϵ_x 和 ϵ_g 是预定的精度值。在实际数值计算中，函数 $g(x)$ 和 $g'(x)$ 只有一定的计算精度，所以算法中对 ϵ_x 和 ϵ_g 的选取并不是越小越好，一般够用就可以

了，取 ϵ 太小有可能导致算法在 x^* 附近反复跳跃而不能满足收敛法则。判断算法收敛，可以使用绝对变化量也可以使用相对变化量。为了保险起见，还应该设置一个迭代最大次数，迭代超过最大次数时宣告算法失败。

牛顿法是方程求根方法, 并通过求导数 $g(x)$ 的根来求最小值点。这里给出牛顿法的一个几何解释。设 x_t 是根 x^* 附近的一个点，从曲线 $g(x)$ 上的点 $(x_t, g(x_t))$ 作切线 (见图6.2)，切线方程为

$$y = g(x_t) + g'(x_t)(x - x_t),$$

用切线在 x_t 附近作为曲线 $g(x)$ 的近似，把求 $g(x) = 0$ 的根近似地变成求切线与 x 轴交点的问题，令

$$0 = g(x_t) + g'(x_t)(x - x_t),$$

则交点 x 的值就是公式(6.30)的下一个迭代值 x_{t+1}。

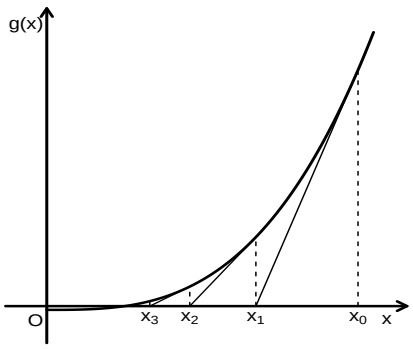

图 6.2: 用牛顿法解一元非线性方程

例 6.2.2. 对函数 $f(x) = \frac{1}{4}x^4 - \frac{1}{8}x$，求 $\arg\min_{x\geq 0} f(x)$。令 $g(x) = f'(x) = x^3 - \frac{1}{8}$，显然 $g(x) = 0$ 在 $x \geq 0$ 有唯一的根 $x^* = \frac{1}{2}$。取 $x_0 = 2$，用(6.30)进行迭代的过程如图6.2。迭代的前几个近似值为 $x_1 = 1.3438$, $x_2 = 0.9189$, $x_3 = 0.6619$。从图中可以看出，牛顿法当 $g(x)$ 斜率大且形状接近直线时收敛最快。 □

如果待求根的函数 $g(x)$ 的导数 $g'(x)$ 没有表达式或很难推导，也可以用数值微分方法计算 $g'(x)$。另一种方法是用函数图像上连接两次迭代的两个点 $(x_{t-1}, g(x_{t-1}))$ 和 $(x_t, g(x_t))$ 的连线斜率代替公式(6.30)中的 $g'(x_t)$，公式变成

$$x_{t+1} = x_t - \frac{x_t - x_{t-1}}{g(x_t) - g(x_{t-1})} g(x_t), \tag{6.32}$$

这种方法称为**割线法**，适当正则条件下具有超线性收敛速度但比牛顿法差一些。割线法简单易行，但是和牛顿法一样，遇到 $g'(x)$ 接近于零的点会使得近似点被抛离真值 x^*。有一些方法结合割线法与二分法的优点，具有超线性收敛速度而且保证根总在迭代的区间内，参见 Monahan(2001)[31] §8.3。

6.2.3 一维搜索的区间 *

一维搜索一般需要先确定包含极小值点的区间, 方程求根时可能也需要确定根所在的区间。

设定义于 $(-\infty,\infty)$ 或 $(0,\infty)$ 的连续目标函数 $f(x)$ 存在局部极小值点 x^*, 且设 $f(x)$ 在 x^* 的一个邻域内是凸函数, 只要能找到三个点 $a<c<b$ 使得 $f(a)>f(c)$, $f(c)<f(b)$，则在 $[a,b]$ 内必存在 $f(x)$ 的一个极小值点。为了求 a, b, 从两个初始点 x_0,x_1 出发，如果 $f(x_0)>f(x_1)$，则最小值点在 x_1 右侧，向右侧搜索找到一个函数值超过 $f(x_1)$ 的点；如果 $f(x_0)<f(x_1)$，则最小值点在 x_0 左侧，向左侧搜索 (参考图6.3)。

求区间 $[a,b]$ 的算法如下:

取步长 $h>0$，步长倍数 $\gamma>1$, 初始点 x_0，$x_1\leftarrow x_0+h$, $k\leftarrow 0$
if $(f(x_0)>f(x_1))$ { # 这时可以向右搜索得到 a,c,b
    until $(f(x_{k+1})>f(x_k))$ {
        $k\leftarrow k+1$
        $x_{k+1}\leftarrow x_k+\gamma^k h$
    }
    $a\leftarrow x_{k-1}, c\leftarrow x_k, b\leftarrow x_{k+1}$
} **else** { # 这时可以向左搜索
    交换 x_0 与 x_1，这样 $x_1<x_0$, $f(x_0)>f(x_1)$
    until $(f(x_{k+1})>f(x_k))$ {
        $k\leftarrow k+1$
        $x_{k+1}\leftarrow x_k-\gamma^k h$
    }
    $a\leftarrow x_{k+1}, c\leftarrow x_k, b\leftarrow x_{k-1}$
}
输出 $[a,b]$ 作为包含极小值点的区间

如果函数 $f(x)$ 的定义域为 $x\in(0,\infty)$，以上的算法应取初值 $x_0>0$，如果出现向左搜索，可取迭代公式为 $x_{k+1}=\frac{1}{\gamma}x_k$。

如果目标函数 $f(x)$ 连续可导，则求最小值点所在区间也可以通过求 $a<b$ 使得 $f'(a)<0$, $f'(b)>0$ 来解决。可取初值 x_0，如果 $f'(x_0)<0$，则类似上面算法那样向右搜索找到 b 使得 $f'(b)>0$；如果 $f'(x_0)>0$，则向左搜索找到 $a<x_0$ 使得 $f'(a)<0$。用最后得到的两个点作为 a,b。可以看出，这里描述的搜索方法适用于一元函数求根时寻找包含根的区间的问题。

6.2.4 0.618 法

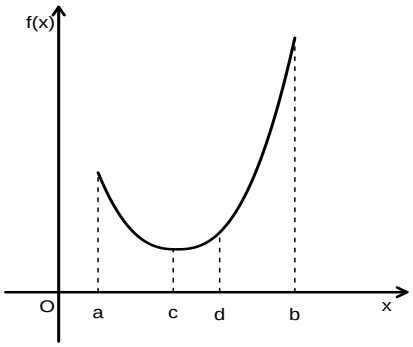

图 6.3: 0.618 法

0.618 法（黄金分割法）是一种常用的不需要导数信息的一维搜索方法。假设 $f(x)$ 在闭区间 $[a,b]$ 内是凸函数，存在最小值点 $x^*\in(a,b)$，这时可以用 0.618 法求最小值点。

在区间 $[a,b]$ 内对称地取两个点 $c<d$ 在所谓的黄金分割点上 (见图6.3)，即

$$\frac{d-a}{b-a}=\rho\triangleq\frac{\sqrt{5}-1}{2}\approx 0.618,\quad \frac{b-c}{b-a}=\rho,$$

黄金分割比例 ρ 满足

$$\frac{1-\rho}{\rho}=\rho,$$

这样的比例的特点是，当区间缩小到 $[a,d]$ 时，c 仍为缩小后的区间的右侧黄金分割点；当区间缩小到 $[c,b]$ 时，d 仍为缩小后的区间的左侧黄金分割点。这样选了 c,d 两个点后，比较 $f(c)$ 和 $f(d)$ 的值，如果 $f(c)$ 较小，则因为 $a<c<d$, $f(a)>f(c)$, $f(c)<f(d)$, 最小值点 x^* 一定在区间 $[a,d]$ 内，可以把搜索区间缩小到 $[a,d]$, 以 $[a,d]$ 作为含有最小值点的区间再比较其中两个黄金分割点的函数值，而且这时 c 是 $[a,d]$ 中右侧的黄金分割点，只要再添加左侧的黄金分割点就可以了。如果比较 $f(c)$ 和 $f(d)$ 时发现 $f(d)$ 较小，则把搜索区间缩小到 $[c,b]$,

这时 d 是 $[c,b]$ 内的左侧黄金分割点，只要再添加右侧黄金分割点。每次迭代使得区间长度缩小到原来的 0.618 倍, 缩小后的两个黄金分割点中一个的坐标和函数值是已经算过的。

设规定的最小值点绝对误差限为 ϵ，算法如下:

令 $\rho \leftarrow 0.618,\ t \leftarrow 0$
令 $a_0 \leftarrow a,\ b_0 \leftarrow b$
令 $c_0 \leftarrow \rho a_0 + (1-\rho)b_0,\ d_0 \leftarrow (1-\rho)a_0 + \rho b_0$
until $(b_t - a_t < \epsilon)$) {
    if($f(c_t) < f(d_t)$) { # 保留左侧
        $a_{t+1} \leftarrow a_t,\ b_{t+1} \leftarrow d_t$
        $d_{t+1} \leftarrow c_{t+1},\ c_{t+1} \leftarrow \rho a_{t+1} + (1-\rho)b_{t+1}$
    } **else** { # 保留右侧
        $a_{t+1} \leftarrow c_t,\ b_{t+1} \leftarrow b_t$
        $c_{t+1} \leftarrow d_t,\ d_{t+1} \leftarrow (1-\rho)a_{t+1} + \rho b_{t+1}$
    }
    $t \leftarrow t+1$
} # end until
令 $x^* \leftarrow \frac{1}{2}(a_t + b_t)$
输出 x^* 作为最小值点近似值

0.618 法具有线性收敛速度。

6.2.5 抛物线法 *

牛顿法需要计算导数，在初值合适的情况下收敛快。牛顿法本质上是利用 $f(x)$ 的一、二阶导数对 $f(x)$ 用二次多项式进行近似并求近似函数的最小值点作为下一个近似值。事实上，为了用二次多项式逼近 $f(x)$，只要有 $f(x)$ 函数图像上的三个点就可以。仍假设 $f(x)$ 在闭区间 $[a,b]$ 内是凸函数，存在最小值点 $x^* \in (a,b)$。设 $c \in (a,b)$, 利用 $(a,f(a)),(c,f(c)),(b,f(b))$ 三个点求得二次插值多项式 $\varphi(x)$，用 $\varphi(x)$ 的最小值点作为下一个近似值。

设插值多项式为

$$\varphi(x) = \beta_0 + \beta_1 x + \beta_2 x^2,$$

其中 $\beta_2 > 0$, 在给定了 $a < c < b$ 和相应函数值后，可以解出 (见习题10)

$$\beta_2 = \frac{1}{b-a}\left[\frac{f(b)-f(c)}{b-c} - \frac{f(c)-f(a)}{c-a}\right], \tag{6.33}$$

$$\beta_1 = \frac{f(c)-f(a)}{c-a} - \beta_2(c+a), \tag{6.34}$$

从而求得 $\varphi(\cdot)$ 的最小值点

$$\tilde{x} = -\frac{\beta_1}{2\beta_2}, \tag{6.35}$$

如果 $\tilde{x} \in (a, c)$, 则以 $a, \tilde{x}, c$ 为新的插值节点继续计算插值多项式并求插值多项式的最小值点；如果 $\tilde{x} \in (c, b)$, 则以 $c, \tilde{x}, b$ 为新的插值节点继续计算插值多项式并求插值多项式的最小值点，如此迭代直至三个插值点足够接近。

适当条件下抛物线法达到超线性收敛速度。

6.2.6 沃尔夫准则 *

在求多元函数优化问题时经常需要进行一维搜索，一维搜索可以在搜索方向上求精确最小值点，也可以求一个使得函数值下降足够多的点，称为不精确一维搜索。

设目标函数 $f(\boldsymbol{x})$ 有连续二阶偏导数，当前的最小值点近似值为 $\boldsymbol{x}^{(t)}$，当前的搜索方向为 $\boldsymbol{d}^{(t)}$, 满足 $[\nabla f(\boldsymbol{x}^{(t)})]^T \boldsymbol{d}^{(t)} < 0$。令 $h(\alpha) = f(\boldsymbol{x}^{(t)} + \alpha \boldsymbol{d}^{(t)})$, $\alpha \geq 0$，则

$$h'(\alpha) = [\nabla f(\boldsymbol{x}^{(t)} + \alpha \boldsymbol{d}^{(t)})]^T \boldsymbol{d}^{(t)}, \tag{6.36}$$

$$h'(0) = [\nabla f(\boldsymbol{x}^{(t)})]^T \boldsymbol{d}^{(t)} < 0, \tag{6.37}$$

$$h(\alpha) = h(0) + h'(0)\alpha + o(\alpha), \tag{6.38}$$

由 $h'(\alpha)$ 的连续性可知 $h(\alpha)$ 在 0 的一个右侧邻域中严格单调下降。

取 $\mu \in (0, \frac{1}{2})$, $\sigma \in (\mu, 1)$，不精确一维搜索的沃尔夫准则要求步长 α 满足如下两个条件:

$$h(0) - h(\alpha) \geq \mu[-h'(0)]\alpha, \tag{6.39}$$

$$-h'(\alpha) \leq \sigma[-h'(0)]. \tag{6.40}$$

第一个限制条件要求函数值下降足够多，第二个限制条件要求 $|h'(\alpha)|$ 已经比较接近于零（精确一维搜索要求 $h'(\alpha) = 0$）。有许多个 α 可以使这两个条件成立。把这两个条件用目标函数 $f(\cdot)$ 表示，可以写成

$$f(\boldsymbol{x}^{(t)}) - f(\boldsymbol{x}^{(t)} + \alpha \boldsymbol{d}^{(t)}) \geq \mu\left\{-[\nabla f(\boldsymbol{x}^{(t)})]^T \boldsymbol{d}^{(t)}\right\}\alpha, \tag{6.41}$$

$$-[\nabla f(\boldsymbol{x}^{(t)} + \alpha \boldsymbol{d}^{(t)})]^T \boldsymbol{d}^{(t)} \leq \sigma\left\{-[\nabla f(\boldsymbol{x}^{(t)})]^T \boldsymbol{d}^{(t)}\right\}. \tag{6.42}$$

σ 取值越小，一维搜索越精确，花费的时间也越长。实际中常取 $\mu = 0.1$, $\sigma \in [0.6, 0.8]$。

下面的算法可以找到满足沃尔夫准则的两个条件的步长 α。想法是，如果条件不满足，就利用现有的 $h(\alpha)$ 的函数值和一阶导数值构造 $h(\alpha)$ 的二阶插值多项式，计算插值多项式的最小值点 $\tilde{\alpha}$ 作为 α 的下一个试验值 (见习题11和习题12)。算法如下:

取定 $\mu \in (0, \frac{1}{2})$, $\sigma \in (\mu, 1)$, 最大步长 $\overline{\alpha}$, 令 $\alpha_0 \leftarrow 0$, $\alpha_2 \leftarrow \overline{\alpha}$
计算 $y_0 \leftarrow h(0)$, $y_0' = h'(0) = [\nabla f(\boldsymbol{x}^{(t)})]^T \boldsymbol{d}^{(t)}$
令 $\alpha_1 \leftarrow \alpha_0$, $y_1 \leftarrow y_0$, $y_1' \leftarrow y_0'$
取 $\alpha \in (0, \alpha_2)$
令 finished $\leftarrow$ FALSE
until (finished) {
    计算 $y \leftarrow h(\alpha)$
    if($y_0 - y \geq -\mu y_0' \alpha$) { # 满足沃尔夫条件 1
        计算 $y' \leftarrow h'(\alpha)$
        if($-y' \leq -\sigma y_0'$) { # 满足沃尔夫条件 1,2
            finished $\leftarrow$ TRUE; break
        } **else** { # 满足条件 1 但不满足条件 2, 延伸搜索
            由 α_1, y_1, y_1' 和 α, y' 计算 $\tilde{\alpha} \leftarrow \alpha_1 - \dfrac{\alpha - \alpha_1}{y' - y_1'} y_1'$
            令 $\alpha_1 \leftarrow \alpha$, $y_1 \leftarrow y$, $y_1' \leftarrow y'$, $\alpha \leftarrow \tilde{\alpha}$
        }
    } **else** { # 不满足沃尔夫条件 1, 收缩搜索
        由 α_1, y_1, y_1' 和 α, y 计算 $\tilde{\alpha} \leftarrow -\dfrac{(\alpha^2 - \alpha_1^2) y_1' - 2\alpha_1 (y - y_1)}{2[y - y_1 - (\alpha - \alpha_1) y_1']}$
        令 $\alpha_2 \leftarrow \alpha$, $\alpha \leftarrow \tilde{\alpha}$
    }
} # until
输出 α 作为步长

在 R 软件中，用 `optimize` 函数求一元函数极值点, 函数结合使用了 0.618 法和抛物线法。

6.3 无约束优化方法

6.3.1 分块松弛法

对多元目标函数 $f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d$, 可以反复地分别延每个坐标轴方向进行一维搜索，称为坐标循环下降法。按这种方法进行引申，可以把 $\boldsymbol{x}$ 的分量分成若干组，每次固定其它分量不变，针对某一组分量进行优化，然后轮换其它组进行优化，这样的方法叫做**分块松弛法**。这种方法常常得到比较简单易行的算法，只不过其收敛速度不一定是最优的。

例 6.3.1 (k 均值聚类). 设在 $\mathbb{R}^d$ 中有 n 个点 $\boldsymbol{x}^{(i)}, i = 1, \ldots, n$，要把这 n 个点按照距离分为 k 个类。设 $C = \{C_1, \ldots, C_k\}$, C_j 是分到第 j 个类中的点的集合，$\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k\}$ 是各类的中心（该类中的点的均值）, 聚类问题就是求 C 和 $\boldsymbol{\mu}$ 使得

$$f(\boldsymbol{\mu}, C) = \sum_{j=1}^{k} \sum_{\boldsymbol{x}^{(i)} \in C_j} \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j\|^2 \tag{6.43}$$

最小。

如果已知 $\boldsymbol{\mu}$，则对每个点计算该点到每个 $\boldsymbol{\mu}_j$ 的距离，把该点归类到距离最近的类中。如果已知 C，则 $\boldsymbol{\mu}_j$ 取第 j 类所有点的平均值。

在 $\boldsymbol{\mu}$ 和 C 都未知的情况下求 $f(\boldsymbol{\mu}, \boldsymbol{C})$ 的最小值点，可以使用分块松弛法。作为初始中心，可以随机地选 k 个不同的 $\boldsymbol{x}^{(i)}$ 点当作初始的 $\boldsymbol{\mu}_j, j = 1, \ldots, k$，但是最好能让这 k 个初始中心相互距离远一些。然后，轮流地，先把所有点按照到类中心的距离分配到 k 个类中，再重新计算类中心 $\{\boldsymbol{\mu}_j\}$, 如此重复直到结果不再变动。这个算法简单有效。 □

6.3.2 最速下降法

在目标函数 $f(\boldsymbol{x})$ 一阶可导时，应利用导数（梯度）的信息，向负梯度方向搜索前进，使得每一步的目标函数值都减小。基本的算法为:

取初值 $\boldsymbol{x}^{(0)} \in \mathbb{R}^d$, 置 $t \leftarrow 0$
until (迭代收敛) {
  求 $\boldsymbol{g}^{(t)} \leftarrow -\nabla f(\boldsymbol{x}^{(t)})$
  求最优步长 $\lambda_t > 0$ 使得 $f(\boldsymbol{x}^{(t)} + \lambda_t \boldsymbol{g}^{(t)}) = \min\limits_{\lambda > 0} f(\boldsymbol{x}^{(t)} + \lambda \boldsymbol{g}^{(t)})$
}
输出 $\boldsymbol{x}^{(t)}$ 作为极小值点

在算法中，“迭代收敛”可以用梯度向量长度小于某个预定值，或者两次迭代间函数值变化小于某个预定值来判断。迭代中需要用某种一维搜索方法求步长 λ_t。

在适当正则性条件下，最速下降法可以收敛到局部极小值点并且具有线性收敛速度。但是，最速下降法连续两次的下降方向 $-\nabla f(\boldsymbol{x}^{(t)})$ 和 $-\nabla f(\boldsymbol{x}^{(t+1)})$ 是正交的，这是因为在第 t 步延 $-\nabla f(\boldsymbol{x}^{(t)})$ 搜索时找到点 $\boldsymbol{x}^{(t+1)}$ 时已经使得函数值在该方向上不再变化，下一个搜索方向如果与该方向不正交就不是下降最快的方向。这样，最速下降法收敛速度比较慢。

例 6.3.2. 求解无约束最优化问题

$$\arg\min_{(x_1,x_2)\in\mathbb{R}^2} 4(x_1-1)^2+(x_2-2)^4. \tag{6.44}$$

显然，函数有全局最小值点 $\boldsymbol{x}^*=(1,2)$。

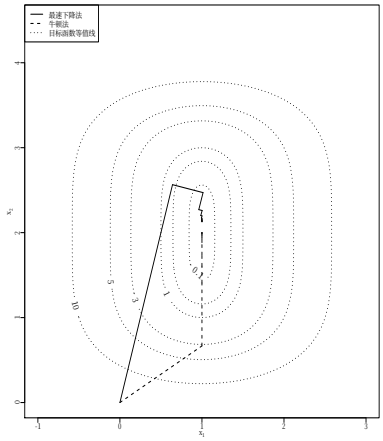

图 6.4: 最速下降法和牛顿法

用最速下降法求解。易见

$$\nabla f(\boldsymbol{x})=\big(8(x_1-1),\ 4(x_2-2)^3\big)^T, \tag{6.45}$$

假设从点 $\boldsymbol{x}^{(0)}=(0,0)$ 出发，函数的等值线图以及迭代轨迹见图6.4中的实线, 轨迹中部分点坐标如下：

$$\boldsymbol{x}^{(1)}=(0.641,2.565),\ \boldsymbol{x}^{(2)}=(1.014,2.471),\ \boldsymbol{x}^{(3)}=(0.962,2.471),\ \boldsymbol{x}^{(4)}=(1.002,2.271),$$
$$\boldsymbol{x}^{(5)}=(0.986,2.202),\ \boldsymbol{x}^{(6)}=(1.001,2.197),\ \ldots,\ \boldsymbol{x}^{(10)}=(1.001,2.156),$$
$$\boldsymbol{x}^{(20)}=(1.001,2.127),\ \boldsymbol{x}^{(40)}=(1.001,2.099),\ \boldsymbol{x}^{(60)}=(1.001,2.083),\ \boldsymbol{x}^{(80)}=(1.001,2.074),$$
$$\boldsymbol{x}^{(100)}=(1.001,2.067),\ \boldsymbol{x}^{(150)}=(1.001,2.055),\ \boldsymbol{x}^{(200)}=(1.001,2.048)$$

□

从例子可以看出，最速下降法的相邻两次搜索方向正交，而且越到最小值点附近接近得越慢。另外，最速下降法需要进行一维搜索。这个方法是其它方法的一个补充，很少单独使用。

6.3.3 牛顿法

牛顿法利用海色阵和梯度向量求下一步，不需要一维搜索，对于二次多项式函数可以一步得到最小值点。对一个存在二阶连续偏导的 d 元函数 $f(\boldsymbol{x})$, 有如下的泰勒近似

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(0)}) + \nabla f(\boldsymbol{x}_0)^T(\boldsymbol{x}-\boldsymbol{x}^{(0)}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^{(0)})^T\nabla^2 f(\boldsymbol{x}^{(0)})(\boldsymbol{x}-\boldsymbol{x}^{(0)}), \tag{6.46}$$

若 $\nabla^2 f(\boldsymbol{x}^{(0)})$ 为正定矩阵，上式右侧的二次多项式函数的最小值点在

$$\boldsymbol{x}^* = \boldsymbol{x}^{(0)} - \left[\nabla^2 f(\boldsymbol{x}^{(0)})\right]^{-1} \nabla f(\boldsymbol{x}^{(0)}) \tag{6.47}$$

处达到。所以牛顿法取初值 $\boldsymbol{x}^{(0)}$ 后，用公式

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \left[\nabla^2 f(\boldsymbol{x}^{(t)})\right]^{-1} \nabla f(\boldsymbol{x}^{(t)}),\ t=0,1,2,\ldots \tag{6.48}$$

进行迭代，直至收敛。收敛的判断准则可以取为 $\|\nabla \boldsymbol{f}(x)\|$ 足够小，或者取为 $|f(\boldsymbol{x}^{(t+1)}) - f(\boldsymbol{x}^{(t)})|$ 足够小。

如果初值接近于最小值点并且目标函数满足一定正则性条件，牛顿法有二阶收敛速度。

例 6.3.3. 再次考虑例6.3.2。易见 $f(\boldsymbol{x})$ 的海色阵为

$$\nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} 8 & 0 \\ 0 & 12(x_2-2)^2 \end{pmatrix},$$

用牛顿法，迭代公式为

$$\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} - \left[\nabla^2 f(\boldsymbol{x}^{(t-1)})\right]^{-1} \nabla f(\boldsymbol{x}^{(t-1)}) = \boldsymbol{x}^{(t-1)} - (x_1-1,\ \frac{1}{3}(x_2-2))^T.$$

迭代过程见图6.4中的虚线。轨迹中部分点坐标如下：

$$\begin{aligned}
&\boldsymbol{x}^{(1)} = (1, 0.667),\ \boldsymbol{x}^{(2)} = (1, 1.111),\ \boldsymbol{x}^{(3)} = (1, 1.407),\ \boldsymbol{x}^{(4)} = (1, 1.605),\\
&\boldsymbol{x}^{(5)} = (1, 1.737),\ \boldsymbol{x}^{(6)} = (1, 1.824),\ \ldots,\ \boldsymbol{x}^{(10)} = (1, 1.965),\\
&\boldsymbol{x}^{(15)} = (1, 1.995),\ \boldsymbol{x}^{(20)} = (1, 1.999)
\end{aligned}$$

没有发生最速下降法那样反复折向前进的问题，而且收敛速度相对较快。 □

6.3.4 拟牛顿法

牛顿法在目标函数具有较好光滑性而且海色阵正定时有很好的收敛速度。但是，牛顿法需要目标函数有二阶导数，还需要在每步迭代都计算海色阵的逆矩阵，在海色阵非正定时可能会失败。

在公式(6.48)中，把其中的海色阵 $\nabla^2 f(\boldsymbol{x}^{(t)})$ 替换成一个保证正定的矩阵 H_t，则 $-H_t^{-1}\nabla f(\boldsymbol{x}^{(t)})$ 是 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}^{(t)}$ 处的一个下降方向，即当 $\lambda>0$ 充分小而且 $\nabla f(\boldsymbol{x}^{(t)})\neq \boldsymbol{0}$ 时一定有

$$f\left(\boldsymbol{x}^{(t)}-\lambda H_t^{-1}\nabla f(\boldsymbol{x}^{(t)})\right)<f(\boldsymbol{x}^{(t)}), \tag{6.49}$$

于是可以用一维搜索方法求步长 λ_{t+1} 使得

$$\begin{cases}\lambda_{t+1}= & \arg\min_{\lambda>0} f\left(\boldsymbol{x}^{(t)}-\lambda H_t^{-1}\nabla f(\boldsymbol{x}^{(t)})\right),\\ \boldsymbol{x}^{(t+1)}= & \boldsymbol{x}^{(t)}-\lambda_t H_t^{-1}\nabla f(\boldsymbol{x}^{(t)}).\end{cases} \tag{6.50}$$

这样得到了保证函数值每次迭代都下降的算法。这里正定矩阵 H_t 的选取有各种各样的方法，比如，取 H_t 为单位阵，则(6.50)变成最速下降法。

在(6.50)中取 H_t 为海色阵 $\nabla^2 f(\boldsymbol{x}^{(t)})$，算法就变成增加了一维搜索的牛顿法，称之为**阻尼牛顿法**。其中的步长 λ_{t+1} 可以用一种一维搜索方法来求，最简单的做法是，依次考虑 $\lambda=1,1/2,1/4,\ldots$，一旦 $f\left(\boldsymbol{x}^{(t)}-\lambda H_t^{-1}\nabla f(\boldsymbol{x}^{(t)})\right)<f(\boldsymbol{x}^{(t)})$ 就停止搜索。阻尼牛顿法也需要海色阵正定，所以实际使用阻尼牛顿法时，如果某步的海色阵不可逆或者 $\left[\nabla f(\boldsymbol{x}^{(t)})\right]^T\left[\nabla^2 f(\boldsymbol{x}^{(t)})\right]^{-1}\nabla f(\boldsymbol{x}^{(t)})\geq 0$, 可以在此步以 $-\nabla f(\boldsymbol{x}^{(t)})$ 为搜索方向 (最速下降法)，这样修改后可以使得阻尼牛顿法每步都减小目标函数值。

最速下降法、牛顿法、阻尼牛顿法在迭代中遇到鞍点 ($\nabla f(\boldsymbol{x}^{(t)})=\boldsymbol{0}$ 但 $\boldsymbol{x}^{(t)}$ 不是极小值点) 时都不能正常运行。这时需要一些特殊的方法求得下降方向。参见徐成贤等 (2002)[10]§3.2.2。

另外一些构造 H_t 的方法可以不用计算二阶偏导数而仅需要计算目标函数值和一阶偏导数值，可以迭代地构造 H_t，这样的算法有很多，统称为**拟牛顿法**或**变尺度法**。这样的 H_t 应该满足如下条件:

(1) H_t 是对称正定阵，从而 $\boldsymbol{d}^{(t)}\triangleq -H_t^{-1}\nabla f(\boldsymbol{x}^{(t)})$ 是下降方向 (称 $\boldsymbol{d}^{(t)}$ 为拟牛顿方向)。

(2) H_{t+1} 由 H_t 迭代得到:

$$H_{t+1}=H_t+\Delta H_t,$$

称 ΔH_t 是修正矩阵。

(3) H_{t+1} 必须满足如下拟牛顿条件:

$$H_{t+1}\boldsymbol{\delta}^{(t)} = \boldsymbol{\zeta}^{(t)}. \tag{6.51}$$

其中 $\boldsymbol{\delta}^{(t)} \triangleq \boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}$, $\boldsymbol{\zeta}^{(t)} \triangleq \nabla f(\boldsymbol{x}^{(t+1)}) - \nabla f(\boldsymbol{x}^{(t)})$。

前两个条件比较自然，下面解释第三个条件的理由。

设迭代已经得到了 H_t 和 $\boldsymbol{x}^{(t+1)}$，需要给出下一步的 H_{t+1}。把 $f(\boldsymbol{x})$ 在 $\boldsymbol{x}^{(t+1)}$ 处作二阶泰勒展开得

$$\begin{aligned} f(\boldsymbol{x}) \approx & f(\boldsymbol{x}^{(t+1)}) + \nabla f(\boldsymbol{x}^{(t+1)})^T(\boldsymbol{x} - \boldsymbol{x}^{(t+1)}) \\ & + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(t+1)})^T \nabla^2 f(\boldsymbol{x}^{(t+1)})(\boldsymbol{x} - \boldsymbol{x}^{(t+1)}), \end{aligned}$$

这意味着

$$\nabla f(\boldsymbol{x}) \approx \nabla f(\boldsymbol{x}^{(t+1)}) + \nabla^2 f(\boldsymbol{x}^{(t+1)})(\boldsymbol{x} - \boldsymbol{x}^{(t+1)}),$$

在上式中取 $\boldsymbol{x} = \boldsymbol{x}^{(t)}$ 得

$$\nabla f(\boldsymbol{x}^{(t)}) \approx \nabla f(\boldsymbol{x}^{(t+1)}) + \nabla^2 f(\boldsymbol{x}^{(t+1)})(\boldsymbol{x}^{(t)} - \boldsymbol{x}^{(t+1)}),$$

则有

$$\nabla^2 f(\boldsymbol{x}^{(t+1)})(\boldsymbol{x}^{(t+1)} - \boldsymbol{x}^{(t)}) \approx \nabla f(\boldsymbol{x}^{(t+1)}) - \nabla f(\boldsymbol{x}^{(t)}) = \boldsymbol{\zeta}^{(t)},$$

因为 H_{t+1} 是用来代替海色阵的，所以根据上式需要 H_{t+1} 满足拟牛顿条件(6.51)。

修正矩阵 ΔH_t 有多种取法，其中一种比较高效而稳定的算法称为 BFGS 算法 (Broyden-Fletcher-Goldfarb-Shanno)，在多元目标函数无约束优化问题中被广泛应用。BFGS 算法的修正公式为

$$H_{t+1} = H_t + \frac{\boldsymbol{\zeta}^{(t)}(\boldsymbol{\zeta}^{(t)})^T}{(\boldsymbol{\zeta}^{(t)})^T\boldsymbol{\delta}^{(t)}} - \frac{(H_t\boldsymbol{\delta}^{(t)})\,(H_t\boldsymbol{\delta}^{(t)})^T}{(\boldsymbol{\delta}^{(t)})^T H_t \boldsymbol{\delta}^{(t)}}, \tag{6.52}$$

一般取初值 H_0 为单位阵，这样按(6.50)和(6.52)迭代，如果一维搜索采用精确一维搜索或沃尔夫准则，(6.52)中的分母 $(\boldsymbol{\zeta}^{(t)})^T\boldsymbol{\delta}^{(t)}$ 可以保证为正值，使得 $\{H_t\}$ 总是对称正定矩阵，搜索方向 $-H_t^{-1}\nabla f(\boldsymbol{x}^{(t)})$ 总是下降方向。在适当正则性条件下 BFGS 方法具有超线性收敛速度。详见徐成贤等 (2002)[10]§3.3。

在(6.50)中计算 $H_t^{-1}\nabla f(\boldsymbol{x}^{(t)})$ 要解一个线性方程组，需要 $O(d^3)$ 阶的计算量 (d 是 $\boldsymbol{x}$ 的维数)。事实上，在 BFGS 算法中 H_t^{-1} 可以递推计算而不需直接求逆或求解线性方程组。记

$V_t = H_t^{-1}$, 这些逆矩阵有如下递推公式:

$$V_{t+1} = V_t + \left[1 + \frac{(\boldsymbol{\zeta}^{(t)})^T V_t \boldsymbol{\zeta}^{(t)}}{(\boldsymbol{\zeta}^{(t)})^T \boldsymbol{\delta}^{(t)}}\right] \frac{\boldsymbol{\delta}^{(t)}(\boldsymbol{\delta}^{(t)})^T}{(\boldsymbol{\zeta}^{(t)})^T \boldsymbol{\delta}^{(t)}} - \frac{H_t \boldsymbol{\zeta}^{(t)}(\boldsymbol{\delta}^{(t)})^T + \left[H_t \boldsymbol{\zeta}^{(t)}(\boldsymbol{\delta}^{(t)})^T\right]^T}{(\boldsymbol{\zeta}^{(t)})^T \boldsymbol{\delta}^{(t)}} \tag{6.53}$$

6.3.5 Nelder-Mead 方法 *

在导数不存在或很难获得时，求解多元函数无约束最小值经常使用 Nelder-Mead 单纯型方法[32]。**单纯型** (simplex) 指在 $\mathbb{R}^d$ 空间中的 $d+1$ 个点作为顶点的图形构成的凸集，比如，在 $\mathbb{R}^2$ 中给定了三个顶点的一个三角形是一个单纯型，在 $\mathbb{R}^3$ 中给定了四个顶点的一个四面体是一个单纯型。

Nelder-Mead 方法是一种直接搜索算法，不使用导数或数值导数。算法开始时要找到 $d+1$ 个点 $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ 构成一个单纯型（要求这 $d+1$ 个点不能位于一个超平面内），计算相应的目标函数值 $y_j = f(\boldsymbol{x}_j), j = 0, 1, \ldots, d$。然后，进行反复迭代，对当前的单纯型进行变换使得目标函数值变小。迭代在单纯型变得足够小或者函数值的变化变得足够小时结束，保险起见，如果迭代次数超过一个预定次数就认为算法失败而结束。单纯型可以沿着目标函数下降的方向延伸，在遇到一个峡谷时可以改变方向，在最小值点附近可以收缩。

选取初值时，可以任意选取初值 $\boldsymbol{x}_0$，然后其他 d 个点可以从 $\boldsymbol{x}_0$ 出发分别延 d 个正坐标轴方向前进一段距离得到。初始的单纯型不要取得太小，否则很容易停留在附近的局部极小值点。

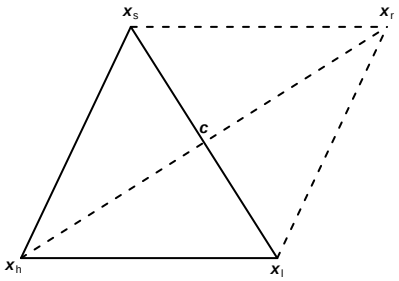

图 6.5: Nelder-Mead 算法反射变换图示，其它变换类似

在算法迭代的每一步,假设当前单纯型为 $(\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$, 相应的函数值为 $(y_0, y_1, \ldots, y_d)$, 首先从中找到最坏的、次坏的、最好的函数值 y_h, y_s, y_l，如果每次得到新的单纯型都把单纯型 $d+1$ 个点次序重排使其函数值由小到大排列，则 $y_l = y_0, y_s = y_{d-1}, y_h = y_d$。然后，计算除最坏点之外的 d 个顶点的重心 $\boldsymbol{c} = \frac{1}{d}\sum_{j \neq h} \boldsymbol{x}_j$。然后对单纯型进行变换以减小函数值。

单纯型的变换有反射、延伸 (expansion)、退缩 (contraction) 和缩小 (shrinkage) 等四种，变化量分别由四个参数 $\alpha > 0, \beta \in (0,1), \gamma > \alpha, \delta \in (0,1)$ 控制，这四个参数一般取为 $\alpha = 1, \beta = \frac{1}{2}, \gamma = 2, \delta = \frac{1}{2}$。

首先考虑反射是否有效，计算反射点 $\boldsymbol{x}_r \triangleq \boldsymbol{c} + \alpha(\boldsymbol{c} - \boldsymbol{x}_h)$, 这实际是把最坏点 $\boldsymbol{x}_h$ 与中心 $\boldsymbol{c}$ 连线后把连线延长得到点 $\boldsymbol{x}_r$(见图6.5)，这样点 $\boldsymbol{x}_h, \boldsymbol{c}, \boldsymbol{x}_r$ 在一条直线上但是最坏点 $\boldsymbol{x}_h$ 与反射点 $\boldsymbol{x}_r$ 分别在 $\boldsymbol{c}$ 的两侧。这样延最坏点的反方向搜索期望能改善目标函数值。如果这时 $y_r \triangleq f(\boldsymbol{x}_r)$ 满足 $y_l \le y_r < y_s$ （反射点比原次坏点好、但不优于原最好点），则接受 $\boldsymbol{x}_r$ 为本迭代步的新顶点，替换原来的最坏点 $\boldsymbol{x}_h$ 即可。

如果反射点 $\boldsymbol{x}_r$ 的目标函数值比原最好点还小 $(y_r < y_l)$, 则考虑把生成 $\boldsymbol{x}_r$ 的延长线进一步延伸，得到延伸点 $\boldsymbol{x}_e \triangleq \boldsymbol{c} + \gamma(\boldsymbol{x}_r - \boldsymbol{c})$。令 $y_e = f(\boldsymbol{x}_e)$，如果比反射点进一步改善 $(y_e < y_r)$, 则接受延伸点 $\boldsymbol{x}_e$ 为本迭代步的新顶点，在单纯型中替换原来的最坏点 $\boldsymbol{x}_h$ 即可，本步迭代结束。如果延伸后 $\boldsymbol{x}_e$ 的函数值不如反射点 $\boldsymbol{x}_r$ 的函数值 $(y_e \ge y_r)$, 则接受反射点 $\boldsymbol{x}_r$ 为本迭代步的新顶点，在单纯型中替换原来的最坏点 $\boldsymbol{x}_h$ 即可，本步迭代结束。

在尝试反射 $\boldsymbol{x}_r$ 后如果发现其不优于次坏点 $\boldsymbol{x}_s (y_r \ge y_s)$, 则反射点不能使用。这时，可以在 $\boldsymbol{x}_h, \boldsymbol{c}, \boldsymbol{x}_r$ 构成的线段上另找一个点，看这个点是否可以接受，这样的变换称为退缩。新的点可以在 $\boldsymbol{c}$ 到 $\boldsymbol{x}_r$ 之间（单纯型外部），也可以在 $\boldsymbol{x}_h$ 和 $\boldsymbol{c}$ 之间（单纯型内部）。这时，如果反射点的函数值 y_r 介于原次坏点与最坏点之间 $(y_s \le y_r \le y_h)$, 那么单纯型外部有希望改善目标函数，所以取 $\boldsymbol{x}_c = \boldsymbol{c} + \beta(\boldsymbol{x}_r - \boldsymbol{c})$(在单纯型外部)；否则，若 y_r 比原最坏点还差 $(y_r > y_h)$, 则只能在单纯型内部考虑，取 $\boldsymbol{x}_c = \boldsymbol{c} - \beta(\boldsymbol{c} - \boldsymbol{x}_h)$（在单纯型内部）。得到 $\boldsymbol{x}_c$ 后令 $y_c = f(\boldsymbol{x}_c)$, 如果比反射点有改善 $(y_c < y_r)$, 则接受退缩点 $\boldsymbol{x}_c$ 为本迭代步的新顶点，在单纯型中替换原来的最坏点 $\boldsymbol{x}_h$ 即可，本步迭代结束。否则，就要执行第四种变换：缩小。

当反射、延伸、退缩都失败时，执行缩小变换。缩小变换是保持原来的最好点 $\boldsymbol{x}_l$ 不动，其它点都按比例向 $\boldsymbol{x}_l$ 收缩: $\boldsymbol{x}_j \leftarrow \boldsymbol{x}_l + \delta(\boldsymbol{x}_j - \boldsymbol{x}_l), j \ne l$。单纯型缩小后希望其它 d 个点的目标函数值能有所改善。

Nelder-Mead 方法的收敛性很难确定。

在 R 软件中，用 `optim` 函数求解多元函数无约束优化问题，该函数提供了 BFGS、Nelder-Mead、共轭梯度、模拟淬火等算法。

6.4 约束优化方法 *

考虑约束最优化问题。约束最优化问题比无约束优化问题复杂得多，也没有很好的通用解决方法，在这个方面有大量研究。现有的方法大致上有如下三类：

- 把约束问题转化为一系列无约束问题，用这些无约束问题的极小值点去逼近约束极小值点，称这样的方法为序列无约束优化方法 (SUMT)，如惩罚函数法、乘子罚函数法。
- 在迭代的每一步用一个二次函数逼近目标函数，用线性约束近似一般约束，构造一系列二次规划来逼近原问题，称这样的方法为序列二次规划 (SQP) 法。
- 每次迭代找到可行下降方向，保证迭代始终处于可行域内，这样的方法称为可行方向法。

6.4.1 约束的化简

最简单的一些约束可以通过参数变换转化成无约束问题。例如，对 $f(x)$ 在 $x \in (0,\infty)$ 求最小值，可令 $x = e^t$, $h(u) = f(e^u), u \in (-\infty,\infty)$，若求得 $\arg\min_{u\in\mathbb{R}} h(u) = u^*$，则 $x^* \triangleq e^{u^*}$ 是 $f(x)$ 在 $(0,\infty)$ 的最小值点。类似地，对 x 限制在有限开区间 (a,b) 的情形，可以作变换

$$x = a + \frac{b-a}{1+e^{-u}},\ u \in (-\infty,\infty),$$

对 x 限制在有限闭区间 $[a,b]$ 的情形，可以做变换

$$x = a + (b-a)\sin^2 u,\ u \in (-\infty,\infty),$$

对 $\boldsymbol{x} = (x_1,x_2,x_3)$ 限制为 $0 \le x_1 \le x_2 \le x_3$ 的情形，可以做变换

$$x_1 = u_1^2,\quad x_2 = u_1^2 + u_2^2,\quad x_3 = u_1^2 + u_2^2 + u_3^2,\ (u_1,u_2,u_3) \in \mathbb{R}^3,$$

对于 $\boldsymbol{x} = (x_1,x_2)$ 限制为 $x_1 > 0, x_2 > 0, x_1 + x_2 < 1$ 的情形，可以做变换

$$x_1 = \frac{e^{t_1}}{1+e^{t_1}+e^{t_2}},\quad x_2 = \frac{e^{t_2}}{1+e^{t_1}+e^{t_2}},\ (t_1,t_2) \in \mathbb{R}^2.$$

这些变换可以把原来的约束优化问题转化为无约束优化问题，或者减少约束个数。

对于等式约束，如果有显式解，可以从中解出部分自变量，简化原来的问题，比如下面的情况。

6.4.2 仅含线性等式约束的情形

如果约束仅包含线性方程组，可以把问题(6.4)用矩阵形式写成

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}),\ \text{s.t.} \\ A^T\boldsymbol{x} = \boldsymbol{b}, \end{cases} \tag{6.54}$$

其中 A 为 $d\times p$ 矩阵, 各列为 $\boldsymbol{a}_i, i=1,\ldots,p$，$\boldsymbol{b}=(b_1,\ldots,b_p)^T$, 即问题仅有 p 个线性约束 $\boldsymbol{a}_i^T\boldsymbol{x}=b_i, i=1,2,\ldots,p$。设 $p<d$ 且 A 列满秩，当 A 规模很小时，可以预先用消元法把 $\boldsymbol{x}$ 的 p 个分量用其余的 $d-p$ 个线性表出，把问题转化为 $d-p$ 为的无约束优化问题。

下面讨论当 A 规模较大时的处理方法。

记 $\mu(A)$ 为 A 的各列在 $\mathbb{R}^d$ 中张成的线性空间, 即

$$\mu(A)\triangleq\{\boldsymbol{x}\in\mathbb{R}^d:\boldsymbol{x}=A\boldsymbol{u},\boldsymbol{u}\in\mathbb{R}^p\}=\{\boldsymbol{x}\in\mathbb{R}^d:\boldsymbol{x}=\sum_{i=1}^{p}c_i\boldsymbol{a}_i,\ c_1,\ldots,c_p\in\mathbb{R}\}\tag{6.55}$$

则 $\mathbb{R}^d$ 可以分解为 $\mu(A)$ 与 $\mu(A)$ 的正交补空间

$$\mu(A)^{\perp}\triangleq\{\boldsymbol{x}\in\mathbb{R}^d:A^T\boldsymbol{x}=0\}\tag{6.56}$$

的直和:

$$\mathbb{R}^d=\mu(A)\oplus\mu(A)^{\perp}.$$

设 A 有 QR 分解

$$A=Q\begin{pmatrix}R\\0\end{pmatrix}=(Q_1,Q_2)\begin{pmatrix}R\\0\end{pmatrix}=Q_1R,$$

其中 Q 为 $d\times d$ 正交阵，R 为 $p\times p$ 上三角阵，Q_1 为 $d\times p$ 矩阵, 易见 Q_1 各列构成 $\mu(A)$ 的标准正交基，Q_2 各列构成 $\mu(A)^{\perp}$ 的标准正交基。若 $\boldsymbol{x}$ 满足约束 $A\boldsymbol{x}=b$, 设

$$\boldsymbol{x}=Q_1\boldsymbol{u}\oplus Q_2\boldsymbol{v},\ \boldsymbol{u}\in\mathbb{R}^p,\ \boldsymbol{v}\in\mathbb{R}^{d-p},$$

则

$$A^T\boldsymbol{x}=R^TQ_1^T\boldsymbol{x}=R^T\boldsymbol{u}=\boldsymbol{b}$$

可得 $\boldsymbol{u}=R^{-T}\boldsymbol{b}$(简记 $(R^T)^{-1}$ 为 R^{-T})，于是

$$\boldsymbol{x}=Q_1R^{-T}\boldsymbol{b}+Q_2\boldsymbol{v},\ \boldsymbol{v}\in\mathbb{R}^{d-p},\tag{6.57}$$

只要对目标函数 $f(Q_1R^{-T}\boldsymbol{b}+Q_2\boldsymbol{v}),\boldsymbol{v}\in\mathbb{R}^{d-p}$ 求解无约束最优化问题即可。

在 $\boldsymbol{x}$ 的分解式(6.57)中，易见

$$\begin{aligned}P\triangleq&Q_2Q_2^T=I_n-Q_1Q_1^T=I_n-AR^{-1}R^{-T}A^T=I_n-A(R^TR)^{-1}A^T\\=&I_n-A(A^TA)^{-1}A^T,\end{aligned}$$

这是向正交补空间 $\mu(A)^{\perp}$ 投影的投影矩阵，所以(6.57)中属于正交补空间 $\mu(A)^{\perp}$ 的部分 $Q_2\boldsymbol{v}$，在已知 $\boldsymbol{x}$ 时可以用投影表示为 $P\boldsymbol{x}$。后文中的投影梯度法利用了这样的思想。

如果 $\boldsymbol{x}^{(t)}$ 是一个可行点，为了使得 $\boldsymbol{x}^{(t+1)}=\boldsymbol{x}^{(t)}+\boldsymbol{d}$ 也是可行点，需要 $A^T\boldsymbol{d}=0$。注意问题(6.54)是一般约束问题(6.4)仅含等式约束的版本，这里 $c_i(\boldsymbol{x})=\boldsymbol{a}_i^T\boldsymbol{x}$, $\boldsymbol{a}_i$ 是 A 的第 i 列，可见 $\nabla c_i(\boldsymbol{x})=\boldsymbol{a}_i$，上述对 $\boldsymbol{d}$ 的要求 $A^T\boldsymbol{d}=0$ 可以表示为

$$\boldsymbol{d}^T\nabla c_i(\boldsymbol{x})=0, i=1,\ldots,p,$$

这个要求可以推广到非线性等式约束的情形，即为了 $\boldsymbol{x}^{(t)}+\boldsymbol{d}$ 继续满足等式约束，搜索方向 $\boldsymbol{d}$ 必须和等式约束函数的梯度都正交，否则会使得 $c_i(\boldsymbol{x})$ 的值改变。

6.4.3 线性约束最优化方法

在约束最优化问题中，如果所有约束都是线性函数，这样的问题相对比较容易处理，但是也能体现约束优化问题的困难。如下的约束优化问题称为**线性约束优化问题**:

$$\begin{cases}\arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}),\ \text{s.t.}\\ \boldsymbol{a}_i^T\boldsymbol{x}=b_i,\ i=1,\ldots,p,\\ \boldsymbol{a}_i^T\boldsymbol{x}\geq b_i,\ i=p+1,\ldots,q,\end{cases}\tag{6.58}$$

其中 $\boldsymbol{a}_i\in\mathbb{R}^d, i=1,\ldots,p+q$, $b_i\in\mathbb{R}, i=1,\ldots,p+q$。

前面已经讨论了仅有线性等式约束的情形，下面讨论含有线性不等式约束的情形。

投影梯度法

从(6.57)看出，当约束为线性等式约束时，搜索时应该在和各 $\boldsymbol{a}_i$ 正交的方向上搜索。这提示我们，对于包含不等式约束的优化问题(6.58), 如果 $\boldsymbol{x}^{(t)}$ 是一个可行点，希望找到使得函数值下降的可行方向，只需考虑所有等式约束和起作用的不等式约束。假设 $\boldsymbol{x}^{(t)}$ 处起作用的不等式约束下标为 $p+1,\ldots,p+r$，记

$$\mathcal{B}=\{\boldsymbol{d}\in\mathbb{R}^d:\ \boldsymbol{a}_i^T\boldsymbol{d}=0,\ i=1,\ldots,p+r\}\tag{6.59}$$

设负梯度 $-\nabla f(\boldsymbol{x}^{(t)})$ 投影到 $\mathcal{B}$ 中得到 $\boldsymbol{d}$，如果 $\boldsymbol{d}\neq\boldsymbol{0}$, $\boldsymbol{d}$ 就是一个可行下降方向, 延方向 $\boldsymbol{d}$ 搜索使得函数值下降且保持所有 $p+q$ 个约束成立。有了可行下降方向 $\boldsymbol{d}$ 以后，从 $\boldsymbol{x}^{(t)}$ 出发沿 $\boldsymbol{d}$ 方向进行一维搜索，得到下一个近似极小值点 $\boldsymbol{x}^{(t+1)}$，如此迭代逼近。

如果投影得到的 $\boldsymbol{d}=\boldsymbol{0}$, 必存在 $\lambda_i^{(t)}, i=1,\ldots,p+r$, 满足

$$\sum_{i=1}^{p+r}\lambda_i^{(t)}\boldsymbol{a}_i=\nabla f(\boldsymbol{x}^{(t)}).$$

解出 $\lambda_i^{(t)}, i=1,\ldots,p+r$, 如果 $\lambda_i^{(t)}\geqslant 0, i=p+1,\ldots,p+r$, 只要令 $\lambda_i^{(t)}=0, i=p+r+1,\ldots,p+q$, $\boldsymbol{\lambda}^{(t)}=(\lambda_1^{(t)},\ldots,\lambda_{p+q}^{(t)})^T$，则 $(\boldsymbol{x}^{(t)},\boldsymbol{\lambda}^{(t)})$ 是问题(6.58)的 KKT 对，算法不再继续，只要设法判断 $\boldsymbol{x}^{(t)}$ 是否极小值点。

如果 $\boldsymbol{d}=\boldsymbol{0}$ 但是解出的 $\lambda_i^{(t)}, i=p+1,\ldots,p+r$ 中有负值，则设 $\lambda_{i_0}^{(t)}=\min\{\lambda_i^{(t)}, i=p+1,\ldots,p+r\}$, 令

$$\mathcal{B}'=\{\boldsymbol{d}\in\mathbb{R}^d:\ \boldsymbol{a}_i^T\boldsymbol{d}=0,\ i=1,\ldots,p+r, i\neq i_0\}$$

令 $\tilde{\boldsymbol{d}}$ 为 $-\nabla f(\boldsymbol{x}^{(t)})$ 向 $\mathcal{B}'$ 的投影，则 $\tilde{\boldsymbol{d}}$ 一定是可行下降方向。

这种方法称为**投影梯度法**，是一种比较早期的可行方向法，具体算法略。投影梯度法以及与之类似的简约梯度法优点是比较简单，缺点是仅利用梯度信息而没有试图采用高阶逼近，收敛速度慢。

有效集方法

有效集方法可以看成是投影梯度法的推广。对问题(6.58)，假设已有约束最小值点的近似值 $\boldsymbol{x}^{(t)}$ 是可行点 (满足约束条件)，设 $\boldsymbol{x}^{(t)}$ 处不等式约束中 $c_i, i=p+1,\ldots,p+r$ 是起作用约束，$c_i, i=p+r+1,\ldots,p+q$ 不起作用。如前所述，搜索方向必须与 $\boldsymbol{a}_i, i=1,\ldots,p+r$ 正交，仍按(6.59)定义 $\mathcal{B}$ 为这些方向的集合。投影梯度法是把负梯度向量投影到线性子空间 $\mathcal{B}$ 上作为搜索方向，有效集方法推广了这种方法，仍要求搜索方向在 $\mathcal{B}$ 中，但不限于负梯度投影方向，即在第 $t+1$ 步求解如下的仅含线性约束的子问题

$$\begin{cases}\arg\min\limits_{\boldsymbol{d}\in\mathbb{R}^d} f(\boldsymbol{x}^{(t)}+\boldsymbol{d}),\ \text{s.t.}\\ \boldsymbol{a}_i^T\boldsymbol{d}=0,\ i=1,\ldots,p+r\end{cases}\tag{6.60}$$

这个子问题可以比较容易地求解。

设解(6.60)得到 $\boldsymbol{d}^{(t)}$。如果 $\boldsymbol{d}^{(t)}=\boldsymbol{0}$，这说明仅含等式约束的子问题

$$\begin{cases}\arg\min\limits_{\boldsymbol{b}\in\mathbb{R}^d} f(\boldsymbol{x}),\ \text{s.t.}\\ \boldsymbol{a}_i^T\boldsymbol{x}=b_i,\ i=1,\ldots,p+r\end{cases}\tag{6.61}$$

以 $\boldsymbol{x}^{(t)}$ 为最小值点，由定理6.1.5，可以解出拉格朗日乘子 $\lambda_i^{(t)}, i=1,\ldots,p+r$ 使得

$$\sum_{i=1}^{p+r}\lambda_i^{(t)}\boldsymbol{a}_i=\nabla f(\boldsymbol{x}).$$

如果 $\lambda_i^{(t)}\geq 0, i=p+1,\ldots,p+r$，只要令 $\lambda_i^{(t)}=0, i=p+r+1,\ldots,p+q$，$\boldsymbol{\lambda}^{(t)}=(\lambda_1^{(t)},\ldots,\lambda_{p+q}^{(t)})^T$，则由定理6.1.6可知 $(\boldsymbol{x}^{(t)},\boldsymbol{\lambda}^{(t)})$ 是问题(6.58)的 KKT 对，只要判断 $\boldsymbol{x}^{(t)}$ 是不是极小值点。

如果 $\boldsymbol{d}^{(t)}=\boldsymbol{0}$ 但是解出的 $\lambda_i^{(t)}, i=p+1,\ldots,p+r$ 中有负值，设 $\lambda_{i_0}^{(t)}=\min\{\lambda_i^{(t)}, i=p+1,\ldots,p+r\}$，只要从子问题(6.60)的约束中删去第 i_0 个重新求解。

如果子问题(6.60)的解 $\boldsymbol{d}^{(t)}\neq\boldsymbol{0}$，这时如果 $\boldsymbol{x}^{(t)}+\boldsymbol{d}^{(t)}$ 是可行点，满足所有的等式约束和不等式约束（包括不起作用的不等式约束），则令 $\boldsymbol{x}^{(t+1)}=\boldsymbol{x}^{(t)}+\boldsymbol{d}^{(t)}$，否则就把 $\boldsymbol{d}$ 当作一个可行下降方向做一维搜索，一维搜索时要注意不起作用约束也要满足。

具体算法从略，详见徐成贤等 (2002)[10]§6.3.1。

6.4.4 二次规划问题

在仅含线性约束的非线性规划问题中，如果 $f(\boldsymbol{x})$ 是二次多项式函数，称这样的问题为**二次规划**问题，这是最简单的非线性规划问题，这种问题有很多实际应用，另外，非线性约束优化问题的求解也往往需要通过求解一系列二次规划子问题实现。

仅含等式约束的二次规划问题

考虑如下的二次规划问题:

$$\begin{cases}\arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} q(\boldsymbol{x})\triangleq\frac{1}{2}\boldsymbol{x}^TH\boldsymbol{x}+\boldsymbol{g}^T\boldsymbol{x}, \quad \text{s.t.}\\ A^T\boldsymbol{x}=\boldsymbol{b},\end{cases}\tag{6.62}$$

其中 H 为 d 阶实对称矩阵，$\boldsymbol{g}\in\mathbb{R}^d$，$A$ 为 $n\times p$ 列满秩矩阵，各列为 $\boldsymbol{a}_i, i=1,\ldots,p$，$\boldsymbol{b}\in\mathbb{R}^p$。此问题可以用6.4.2的方法求解。

设 A 有如下 QR 分解

$$A=Q\begin{pmatrix}R\\0\end{pmatrix}=(Q_1,Q_2)\begin{pmatrix}R\\0\end{pmatrix}=Q_1R,\tag{6.63}$$

其中 Q 为 $d\times d$ 正交阵，R 为 $p\times p$ 满秩上三角阵，Q_1 为 $d\times p$ 矩阵, $Q_1^TQ_1=I_p$, $Q_1^TQ_2=\boldsymbol{0}$。满足约束的可行点 $\boldsymbol{x}$ 的通解为

$$\boldsymbol{x}=Q_1R^{-T}\boldsymbol{b}+Q_2\boldsymbol{v},\quad \boldsymbol{v}\in\mathbb{R}^{d-p},\tag{6.64}$$

令

$$\tilde{q}(\boldsymbol{v}) = q(Q_1 R^{-T}\boldsymbol{b} + Q_2\boldsymbol{v}) \triangleq \frac{1}{2}\boldsymbol{v}^T\tilde{H}\boldsymbol{v} + \tilde{\boldsymbol{g}}^T\boldsymbol{v} + \tilde{c}, \quad \boldsymbol{v} \in \mathbb{R}^{d-p}, \tag{6.65}$$

其中

$$\tilde{H} = Q_2^T H Q_2, \quad \tilde{\boldsymbol{g}} = Q_2^T\boldsymbol{g} + Q_2^T H Q_1 R^{-T}\boldsymbol{b}, \tag{6.66}$$

$\tilde{c}$ 为与 $\boldsymbol{v}$ 无关的常数项。只要求解无约束优化问题 $\arg\min \tilde{q}(\boldsymbol{v})$ 得到 $\boldsymbol{v}^*$，再用(6.64)就可以得到问题(6.62)的最小值点 $\boldsymbol{x}^*$。

在(6.65)的目标函数 $\tilde{q}(\boldsymbol{v})$ 中，海色阵 $\tilde{H}$ 如果有负特征值，则 $\tilde{q}(\boldsymbol{v})$ 的最小值是 $-\infty$，问题无解。事实上，设有 $\lambda < 0$ 以及 $\boldsymbol{d} \neq \boldsymbol{0}$ 使得 $\tilde{H}\boldsymbol{d} = \lambda\boldsymbol{d}$，取 $\boldsymbol{v}^{(k)} = k\boldsymbol{d}$, 则

$$\tilde{q}(\boldsymbol{v}^{(k)}) = \frac{1}{2}\lambda\|\boldsymbol{d}\|^2 \cdot k^2 + \tilde{\boldsymbol{g}}^T\boldsymbol{d} \cdot k \to -\infty, \quad \text{当} k \to \infty. \tag{6.67}$$

所以，不妨设 $\tilde{H}$ 为非负定阵。

如果 $\tilde{H}$ 是正定阵，则 $\tilde{q}(\boldsymbol{v})$ 是严格凸函数，有全局严格最小值点 $\boldsymbol{v}^* = -\tilde{H}^{-1}\tilde{\boldsymbol{g}}$，代入(6.64)可得问题(6.62)的最小值点 $\boldsymbol{x}^*$。这时，再考虑拉格朗日函数

$$L(\boldsymbol{x}, \boldsymbol{\lambda}) = q(\boldsymbol{x}) - \boldsymbol{\lambda}^T(A^T\boldsymbol{x} - \boldsymbol{b}), \tag{6.68}$$

来求拉格朗日函数的稳定点，其中 $\boldsymbol{x}^*$ 已知，由定理6.1.5知拉格朗日乘子 $\boldsymbol{\lambda}^*$ 满足

$$\begin{aligned} \nabla q(\boldsymbol{x}^*) - A\boldsymbol{\lambda}^* &= 0, \\ A\boldsymbol{\lambda}^* &= H\boldsymbol{x}^* + \boldsymbol{g}, \end{aligned}$$

由 $A = Q_1 R$ 得

$$R\boldsymbol{\lambda}^* = Q_1^T(H\boldsymbol{x}^* + \boldsymbol{g}), \tag{6.69}$$

用回代法解此方程可得拉格朗日乘子 $\boldsymbol{\lambda}^*$。

如果 $\tilde{H}$ 仅为非负定阵而不是正定阵，$\tilde{q}(\boldsymbol{v})$ 仍然是凸函数，最小值点与稳定点等价，稳定点存在当且仅当 $\tilde{H}\boldsymbol{v} = -\tilde{\boldsymbol{g}}$ 有解，这等价于 $\tilde{\boldsymbol{g}}$ 属于 $\mu(\tilde{H})$ （$\mu(\tilde{H})$ 是 $\tilde{H}$ 的各列张成的线性子空间），又等价于

$$(I - \tilde{H}\tilde{H}^+)\tilde{\boldsymbol{g}} = 0, \tag{6.70}$$

其中 $\tilde{H}^+$ 是 $\tilde{H}$ 的加号逆。由定理5.6.4，当(6.70)成立时 $\tilde{q}(\boldsymbol{v})$ 有无穷多个最小值点，可以表达为

$$\boldsymbol{v}^* = -\tilde{H}^+\tilde{g} + (I - \tilde{H}^+\tilde{H})\boldsymbol{u}, \quad \forall \boldsymbol{u} \in \mathbb{R}^{n-p}, \tag{6.71}$$

再代入(6.64)可得问题(6.62)的无穷多个最小值点 $\boldsymbol{x}^*$。

在问题(6.62)规模很大的时候，求解线性方程组计算量很大，可以用共轭梯度法迭代地求解，由于二次目标函数 $q(\boldsymbol{x})$ 的特殊性，迭代最多只需要 $n-p$ 次就可以收敛到最小值点。参见徐成贤等 (2002)[10] §6.4.1。

含有不等式约束的二次规划问题

考虑如下的二次规划问题:

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} q(\boldsymbol{x}) \triangleq \frac{1}{2}\boldsymbol{x}^T H\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x}, \quad \text{s.t.} \\ \boldsymbol{a}_i^T\boldsymbol{x} = b_i,\ i = 1,\ldots,p \\ \boldsymbol{a}_i^T\boldsymbol{x} \geq b_i,\ i = p+1,\ldots,p+q \end{cases} \tag{6.72}$$

其中 H 为 d 阶实对称矩阵，$\boldsymbol{g}\in\mathbb{R}^d$, $\boldsymbol{a}_i\in\mathbb{R}^d, i=1,\ldots,p+q$, $b_i\in\mathbb{R}, i=1,\ldots,p+q$。

用前面所述关于一般线性约束的有效集方法可以把含有不等式约束的二次规划问题转化为一系列仅含等式约束的二次规划问题。假设经过 t 步迭代得到问题(6.72)的一个可行点 $\boldsymbol{x}^{(t)}$，$\boldsymbol{x}^{(t)}$ 处不等式约束中 $c_i, i=p+1,\ldots,p+r$ 是起作用约束，$c_i, i=p+r+1,\ldots,p+q$ 不起作用。考虑如下只有等式约束的子问题

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}\boldsymbol{x}^T H\boldsymbol{x} + \boldsymbol{g}^T\boldsymbol{x}, \quad \text{s.t.} \\ \boldsymbol{a}_i^T\boldsymbol{x} = b_i,\ i = 1,\ldots,p+r \end{cases} \tag{6.73}$$

令 $\boldsymbol{x} = \boldsymbol{x}^{(t)} + \boldsymbol{d}$, 问题(6.73)可以等价地写成

$$\begin{cases} \arg\min\limits_{\boldsymbol{d}\in\mathbb{R}^d} \frac{1}{2}\boldsymbol{d}^T H\boldsymbol{d} + (\boldsymbol{g} + H\boldsymbol{x}^{(t)})^T\boldsymbol{d}, \quad \text{s.t.} \\ \boldsymbol{a}_i^T\boldsymbol{d} = 0,\ i = 1,\ldots,p+r \end{cases} \tag{6.74}$$

记 $\mathcal{B} = \{\boldsymbol{d}\in\mathbb{R}^d : \boldsymbol{a}_i^T\boldsymbol{d} = 0, i=1,\ldots,p+r\}$，如果以 $\boldsymbol{a}_i, i=1,\ldots,p+r$ 为列向量组成矩阵 A, 则 $\mathcal{B} = \mu(A)^\perp$。

通过前面对仅含等式约束的二次规划问题的讨论可以知道，如果

$$\boldsymbol{d}^T H\boldsymbol{d} > 0,\ \forall \boldsymbol{d}\in\mathcal{B}\backslash\{\boldsymbol{0}\}, \tag{6.75}$$

则子问题(6.74)存在唯一的全局严格最小值点，设其为 $\boldsymbol{d}^{(t)}$。如果 $\boldsymbol{d}^{(t)} = \boldsymbol{0}$，就说明 $\boldsymbol{x}^{(t)}$ 已经是子问题(6.73)的约束最小值点, 可以解出相应的拉格朗日乘子 $\lambda_i^{(t)}, i=1,\ldots,p+r$。如果解出的 $\lambda_i^{(t)} \geq 0, i=p+1,\ldots,p+r$, 则令 $\lambda_i^{(t)} = 0, i=p+r+1,\ldots,p+q$, $\boldsymbol{\lambda}^{(t)} = (\lambda_1^{(t)},\ldots,\lambda_{p+q}^{(t)})^T$,

这时 $(\boldsymbol{x}^{(t)}, \boldsymbol{\lambda}^{(t)})$ 为原始问题(6.72)的 KKT 对，由于二次目标函数的特殊性以及(6.75)条件可知 $\boldsymbol{x}^{(t)}$ 是问题(6.72)的约束全局严格最小值点。

如果 $\boldsymbol{d}^{(t)} = \boldsymbol{0}$ 但是 $\lambda_{i_0}^{(t)} = \min\{\lambda_i^{(t)} : i = p+1, \ldots, p+r\} < 0$，把对应于 i_0 的约束从子问题(6.74)中删除，可以证明新的子问题有非零解 $\tilde{\boldsymbol{d}}^{(t)}$，且 $\tilde{\boldsymbol{d}}^{(t)}$ 是原问题(6.72)的可行下降方向。

如果 $\boldsymbol{d}^{(t)} \neq \boldsymbol{0}$, 则 $\boldsymbol{d}^{(t)}$ 是原问题(6.72)的可行下降方向。

设 $\boldsymbol{d}^{(t)}$ 是原问题(6.72)的可行下降方向。若 $\boldsymbol{x}^{(t)} + \boldsymbol{d}^{(t)}$ 是原问题(6.72)的可行点，直接取 $\boldsymbol{x}^{(t+1)}$ 继续迭代即可。如果 $\boldsymbol{x}^{(t)} + \boldsymbol{d}^{(t)}$ 超出了(6.72)的可行域，这一定是不起作用的不等式约束中的某些被突破了，可以从 $\boldsymbol{x}^{(t)}$ 出发沿 $\boldsymbol{d}^{(t)}$ 方向进行线性搜索，由二次目标函数的特点知道线性搜索的最小值点在可行域边界处达到，直接取步长为

$$\alpha_t = \min\left\{ \frac{b_i - \boldsymbol{a}_i^T \boldsymbol{x}^{(t)}}{\boldsymbol{a}_i^T \boldsymbol{d}^{(t)}} : \ p+r+1 \leq i \leq p+q \text{且} \boldsymbol{a}_i^T \boldsymbol{d}^{(t)} < 0 \right\} \tag{6.76}$$

注意到 $\boldsymbol{d}^{(t)}$ 已经是子问题(6.74)的约束最小值点，所以如果得到的 $\alpha_t > 1$ 只要取 $\alpha_t = 1$。如果 $\alpha_t < 1$，设(6.76)的最小值是在第 i_1 个约束处达到的，则 $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} + \alpha_t \boldsymbol{d}^{(t)}$ 比 $\boldsymbol{x}^{(t)}$ 增加了一个起作用的不等式约束 i_1，下一步迭代时子问题(6.74)的约束中需要增加约束 $\boldsymbol{a}_{i_1}^T \boldsymbol{x} = b_{i_1}$。

详细算法从略。

6.4.5 非线性约束优化问题

当约束中含有非线性函数时，问题比无约束和线性约束问题都要复杂得多。这时，很难求得可行的下降方向，而且由于数值误差的因素，对于不等式约束很难判断是否起作用。所以一般的非线性约束问题没有通用的解决方法，更没有可靠的软件。来自于线性约束优化的一些做法仍可以采用，但是算法会很复杂。比较容易实现的方法是各类**罚函数法**，定义一系列增加了惩罚项的目标函数，用一系列无约束优化问题的解逼近约束问题的解，这样的方法又称为**序列无约束最小化方法** (Sequential Unconstrained Minimization Teckniques, SUMT)，特点是实现简单，但是收敛慢而且惩罚很重时问题适定性很差，所以难以求得比较精确的结果。另一类方法是在每步迭代时构造一个二次规划子问题作为近似，通过求解一系列二次规划问题逼近一般非线性约束问题的解，这种方法称为**序列二次规划法** (Sequential Quadratic Programming, SQP)，是现在较好的方法。

外点罚函数法

罚函数法引入包括原来的目标函数和对偏离约束的惩罚两部分的新目标函数，对新目标函数求无约束最小值点。

对一般的约束优化问题(6.4), 等式约束可以用 $|c_i(\boldsymbol{x})|$ 惩罚,不等式约束可以用 $|\min(0, c_i(\boldsymbol{x}))|$ 惩罚，取惩罚函数为

$$\bar{p}(\boldsymbol{x}) = \sum_{i=1}^{p} |c_i(\boldsymbol{x})|^2 + \sum_{i=p+1}^{p+q} |\min(0, c_i(\boldsymbol{x}))|^2, \tag{6.77}$$

当且仅当 $\boldsymbol{x}$ 是可行点时 $\bar{p}(\boldsymbol{x}) = 0$。定义新的目标函数

$$\bar{f}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sigma\bar{p}(\boldsymbol{x}), \tag{6.78}$$

其中 $\sigma > 0$ 称为罚因子。若 $\bar{\boldsymbol{x}}$ 是 $\bar{f}(\boldsymbol{x})$ 的一个无约束极值点且 $\bar{\boldsymbol{x}}$ 是(6.4)的可行点，则对(6.4)的任意可行点 $\tilde{\boldsymbol{x}}$，都有

$$\bar{f}(\bar{\boldsymbol{x}}) = f(\bar{\boldsymbol{x}}) + \sigma\bar{p}(\bar{\boldsymbol{x}}) = \min\left[f(\boldsymbol{x}) + \sigma\bar{p}(\boldsymbol{x})\right] \le f(\tilde{\boldsymbol{x}}) + \sigma\bar{p}(\tilde{\boldsymbol{x}}) = f(\tilde{\boldsymbol{x}}) \tag{6.79}$$

即只要无约束问题 $\arg\min \bar{f}(\boldsymbol{x})$ 的解是约束问题的可行点就是约束极小值点。σ 越大，对无约束极小值点不在可行域内的惩罚越大，所以 $\arg\min \bar{f}(\boldsymbol{x})$ 也越可能落在可行域内。算法迭代进行，只要找到的无约束解不可行就增大罚因子 σ 然后继续求解无约束极值点。算法如下:

取精度 $\epsilon > 0$, 倍数 $c > 1$
取 $t \leftarrow 0$, 初始罚因子 $\sigma_0 > 0$, 任取初始点 $\boldsymbol{x}^{(0)} \in \mathbb{R}^d$
until $(\sigma_t\bar{p}(\boldsymbol{x}^{(t)}) < \epsilon)$ {
    $t \leftarrow t + 1$, $\sigma_t \leftarrow c\sigma_{t-1}$
    以 $\boldsymbol{x}^{(t-1)}$ 为初值求解无约束问题 $\arg\min\left[f(\boldsymbol{x}) + \sigma_t\bar{p}(\boldsymbol{x})\right]$ 得 $\boldsymbol{x}^{(t)}$
}

算法实现时可以取 $\sigma_0 = 1$，$c = 10$。求解无约束问题时迭代停止的条件可以取为梯度向量长度 $\|\nabla\bar{f}(\boldsymbol{x})\|$ 小于预定精度值。

运用这样的算法，可以任意选取初始点，不需要初始点在可行域内，每次迭代得到的近似解都在可行域外，只有最后一个解近似地落入可行域而且是近似约束最小值点，所以这种方法称为**外点罚函数法**。这种方法要求目标函数在 $\mathbb{R}^d$ 上都有定义，如果目标函数仅在与可行域有关的子集上有定义则无法使用；最后的解不一定严格满足可行性条件，如果要求解严格可行此种方法也无法使用。另外，随着 σ 的增大，在新目标函数 $\bar{f}(\boldsymbol{x})$ 中目标函数占的比例越来越小，约束惩罚占的比例越来越大，使得优化问题的适定性变得越来越差，求解无约束优化问题会变得很困难，计算误差变得很大。所以，罚函数法虽然看起来很简单，但是不一定能得到好的结果。

内点罚函数法

内点罚函数法要求每次迭代严格地在可行域内进行。定义新的无约束目标函数，当近似极小值点接近可行域边界时新目标函数变得很大（通常在可行域边界处趋于无穷大），相当于在可行域边界处设立了无限高的墙壁不允许搜索越过可行域边界，所以这种方法又称为**障碍罚函数法**。内点罚函数法不能处理含有等式约束的问题，而且要求可行域含有内点。对于如下仅含不等式约束的优化问题

$$\begin{cases} \arg\min\limits_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}), \text{ s.t.} \\ c_i(\boldsymbol{x}) \geq 0, \ i = 1, \ldots, q \end{cases} \tag{6.80}$$

定义罚函数为 $\bar{p}(\boldsymbol{x}) = -\sum_{i=1}^{q} \log c_i(\boldsymbol{x})$ 或 $\bar{p}(\boldsymbol{x}) = \sum_{i=1}^{q} \frac{1}{c_i(\boldsymbol{x})}$，$\boldsymbol{x}$ 只能是严格可行点，即所有 $c_i(\boldsymbol{x}) > 0$, 当某个 $c_i(\boldsymbol{x}) = 0$ 时 $\bar{p}(\boldsymbol{x}) = +\infty$。取新的目标函数

$$\bar{f}(\boldsymbol{x}) = f(\boldsymbol{x}) + \sigma\bar{p}(\boldsymbol{x}), \tag{6.81}$$

因为真正的约束最小值点允许取边界处的值，需要取 $\{\sigma_t\}$ 使得 $\sigma_t \to 0$, 如此减轻对接近可行域边界的惩罚。内点罚函数法和外点罚函数法类似，先取较大的 σ，求 $\bar{f}(\boldsymbol{x})$ 的无约束最小值点，每次迭代减小 σ 的值后求无约束极值点，直到近似约束最小值点达到足够精度。

由于可行域边界处的高墙的存在，内点罚函数法在近似点靠近可行域边界时加罚的目标函数 $\bar{f}(\boldsymbol{x})$ 也会变得病态，难以求得最小值点，并且求 $\bar{f}(\boldsymbol{x})$ 的最小值点时要注意搜索不能跳出可行域。内点罚函数法要求初值在可行域内，可以设法先找可行点，另一种办法是把内点罚函数和外点罚函数结合使用，对等式约束和初始点不满足的不等式约束，可以采用外点罚函数，其它不等式约束采用内点罚函数，这样的方法称为**混合罚函数法**。

乘子罚函数法针对普通罚函数方法的病态问题做了改进，参见高立 (2014)[3] §7.3、§7.4。

序列二次规划法 (SQP)

序列二次规划法 (SQP 方法) 每一步迭代解决一个二次规划子问题，这种方法在中小规模的非线性约束规划问题中是比较好的方法。SQP 的思想类似于无约束规划中的牛顿法，在局部对目标函数用二次函数逼近，对约束用线性函数逼近，得到二次规划子问题。

对一般约束优化问题(6.4)，设 f 有二阶连续偏导数, 设各 c_i 有一阶连续偏导数。设 $\boldsymbol{x}^{(t)}$ 为迭代 t 步后得到的近似约束最小值点，在 $\boldsymbol{x}^{(t)}$ 处对 $f(\boldsymbol{x})$ 作如下的二阶泰勒展开近似:

$$f(\boldsymbol{x}^{(t)} + \boldsymbol{d}) \approx f(\boldsymbol{x}^{(t)}) + [\nabla f(\boldsymbol{x}^{(t)})]^T \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^T [\nabla^2 f(\boldsymbol{x}^{(t)})]^T \boldsymbol{d}, \ \boldsymbol{d} \in \mathbb{R}^d,$$

对 $c_i(\boldsymbol{x})$ 作如下的一阶泰勒展开近似:

$$c_i(\boldsymbol{x}^{(t)}+\boldsymbol{d}) \approx c_i(\boldsymbol{x}^{(t)}) + [\nabla c_i(\boldsymbol{x}^{(t)})]^T\boldsymbol{d},\ \boldsymbol{d}\in\mathbb{R}^d.$$

进行这样的近似后，考虑如下的二次规划子问题

$$\begin{cases} \underset{\boldsymbol{d}\in\mathbb{R}^d}{\arg\min}\ \frac{1}{2}\boldsymbol{d}^T[\nabla^2 f(\boldsymbol{x}^{(t)})]^T\boldsymbol{d} + [\nabla f(\boldsymbol{x}^{(t)})]^T\boldsymbol{d}, \quad \text{s.t.} \\ c_i(\boldsymbol{x}^{(t)}) + [\nabla c_i(\boldsymbol{x}^{(t)})]^T\boldsymbol{d} = 0, i = 1,\ldots,p, \\ c_i(\boldsymbol{x}^{(t)}) + [\nabla c_i(\boldsymbol{x}^{(t)})]^T\boldsymbol{d} \geq 0, i = p+1,\ldots,p+q \end{cases} \tag{6.82}$$

通过求解这样的二次规划子问题得到下一个近似点 $\boldsymbol{x}^{(t+1)}$。

SQP 算法需要考虑的问题比较多，这里略去详细的讨论，感兴趣的读者可以参考高立 (2014)[3] 第九章。

6.5 统计计算中的优化问题 *

统计中许多问题的计算最终都归结为一个最优化问题，典型代表是最大似然估计 (MLE)、各种拟似然估计方法、非线性回归、惩罚函数方法（如 svm、lasso）等。

在统计优化问题中，目标函数往往不象数学函数优化问题那样可以计算到机器精度，目标函数精度一般都是有限的，甚至于是用随机模拟方法计算的。这样，统计问题优化需要稳定性比较高的算法，对于目标函数的小的扰动应该具有良好适应能力，算法迭代收敛条件也不要取得过于严苛，否则可能无法收敛。比如，在最大似然估计问题中，样本值精度受限于测量精度，实际上都存在较大的测量误差，如果优化算法结果对于样本值的微小变化就产生很大变化，这样的算法就是不可取的。

6.5.1 最大似然估计

最大似然估计经常需要用最优化算法计算，最大似然估计问题有自身的特点，可以直接用一般优化方法进行最大似然估计的计算，但是利用最大似然估计的特点可以得到更有效的算法。

设总体 $\boldsymbol{X}$ 有密度或概率函数 $p(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta}$ 为 m 维的分布参数。有了一组样本 $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ 后，似然函数为

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{X}_i|\boldsymbol{\theta}), \tag{6.83}$$

对数似然函数为

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{X}_i|\boldsymbol{\theta}), \tag{6.84}$$

得分法

设对数似然函数 $l(\boldsymbol{\theta})$ 有二阶连续偏导数, $\nabla l(\boldsymbol{\theta})$ 称为得分函数 (score function), 求最大似然估计可以通过解方程 $\nabla l(\boldsymbol{\theta}) = \mathbf{0}$ 实现，这个方程称为**估计方程**。设参数真值为 $\boldsymbol{\theta}_*$, $l(\boldsymbol{\theta})$ 最大值点为 $\hat{\boldsymbol{\theta}}$, 在适当正则性条件下 $\hat{\boldsymbol{\theta}}$ 渐近正态分布:

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_*)), \ n \to \infty, \tag{6.85}$$

其中

$$I(\boldsymbol{\theta}_*) = \mathrm{Var}(\nabla \log p(X|\boldsymbol{\theta}_*)) = \frac{1}{n}\mathrm{Var}(\nabla l(\boldsymbol{\theta}_*)) = E\left[-\nabla^2 \log p(X|\boldsymbol{\theta}_*)\right] = \frac{1}{n}E\left[-\nabla^2 l(\boldsymbol{\theta}_*)\right] \tag{6.86}$$

是 X 的信息阵。记 $I_n(\boldsymbol{\theta}_*) = \mathrm{Var}(\nabla l(\boldsymbol{\theta}_*))$，称 $I_n(\boldsymbol{\theta}_*)$ 为 $(X_1, X_2, \ldots, X_n)$ 的信息阵, 当 $X_1, X_2, \ldots, X_n$ 独立同分布时 $I_n(\boldsymbol{\theta}_*) = nI(\boldsymbol{\theta}_*)$。

注意最大似然估计求最大值点，与前面求最小值点的问题略有差别，有时讨论负对数似然函数更方便。因为多元正态分布的负对数似然函数是正定二次型，所以如果初值取得比较合适，负对数似然函数 $-l(\boldsymbol{\theta})$ 与多元正态分布的负对数似然函数相近，接近于正定二次型，这时求 $-l(\boldsymbol{\theta})$ 的最小值点会比较容易。求得最大似然估计 $\hat{\boldsymbol{\theta}}$ 后，可以用 $\left[I_n(\hat{\boldsymbol{\theta}})\right]^{-1}$ 估计 $\hat{\boldsymbol{\theta}}$ 的协方差阵。

$l(\boldsymbol{\theta})$ 最大值点 $\hat{\boldsymbol{\theta}}$ 的求解可以用通常的牛顿法、BFGS 法，用到的对数似然函数偏导数和二阶偏导数可以推导解析表达式来计算，也可以用数值微分代替。用数值微分可以避免偏导数推导和编程时的错误。如果用牛顿法求最大值点，$\hat{\boldsymbol{\theta}}$ 的协方差阵可以用 $[-\nabla^2 l(\hat{\boldsymbol{\theta}})]^{-1}$ 来估计。

注意到当样本量充分大且 $\boldsymbol{\theta}$ 接近于真值 $\boldsymbol{\theta}_*$ 时海色阵 $\nabla^2 l(\boldsymbol{\theta}) \approx -I_n(\boldsymbol{\theta})$，如果信息阵的公式很容易得到，在牛顿法的迭代中可以用负信息阵代替对数似然函数的海色阵，迭代为

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \hat{\boldsymbol{\theta}}^{(t)} - \left[-I_n(\hat{\boldsymbol{\theta}}^{(t)})\right]^{-1} \nabla l(\hat{\boldsymbol{\theta}}^{(t)}). \tag{6.87}$$

这种方法叫做**得分法** (scoring)。

例 6.5.1 (逻辑斯谛回归参数估计). 设 $Y_i \sim \mathrm{B}(m_i, \pi_i)$, $i = 1, 2, \ldots, n$, $Y_1, Y_2, \ldots, Y_n$ 相互独立，其中 $\pi_i = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)/[1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)]$, $\boldsymbol{\beta}$ 为未知参数向量，自变量 $\boldsymbol{x}_i, i = 1, 2, \ldots, n$ 已知。这

样的模型称为**逻辑斯谛回归模型**，函数 $\mathrm{logit}(\pi) \triangleq \log \frac{\pi}{1-\pi}$, $\pi \in (0,1)$ 称为逻辑斯谛函数，其反函数为 $\mathrm{logit}^{-1}(\gamma) = \frac{\exp(\gamma)}{1+\exp(\gamma)}$。上面模型中的 π_i 满足 $\mathrm{logit}(\pi_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$, $\pi_i = \mathrm{logit}^{-1}(\boldsymbol{\beta}^T \boldsymbol{x}_i)$。

Y_i 的概率函数及其对数为

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} = \left[1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right]^{-m_i} \left[\exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)\right]^{y_i},$$

$$\log P(Y_i = y_i) = \log \binom{m_i}{y_i} + y_i \cdot \boldsymbol{\beta}^T \boldsymbol{x}_i - m_i \log\left[1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right],$$

对数似然函数为（省略了不随 $\boldsymbol{\beta}$ 变化的部分）

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \cdot \boldsymbol{\beta}^T \boldsymbol{x}_i - m_i \log\left[1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)\right] \right\} \tag{6.88}$$

梯度和海色阵为

$$\nabla l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \boldsymbol{x}_i - \frac{m_i \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)} \boldsymbol{x}_i \right) = \sum_{i=1}^{n} (y_i - m_i \pi_i) \boldsymbol{x}_i, \tag{6.89}$$

$$\nabla^2 l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} m_i \frac{\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)}{[1 + \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)]^2} \cdot \boldsymbol{x}_i \boldsymbol{x}_i^T = -\sum_{i=1}^{n} m_i \pi_i (1-\pi_i) \boldsymbol{x}_i \boldsymbol{x}_i^T. \tag{6.90}$$

因为海色阵不包含随机成分，所以 $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ 的信息阵

$$I_n(\boldsymbol{\beta}) = E\left(-\nabla^2 l(\boldsymbol{\beta})\right) = -\nabla^2 l(\boldsymbol{\beta}), \tag{6.91}$$

得分法和普通牛顿法是相同的, 迭代公式同为

$$\begin{cases} \pi_i^{(t)} = \mathrm{logit}^{-1}\left([\boldsymbol{\beta}^{(t)}]^T \boldsymbol{x}_i \right), \ i = 1, 2, \ldots, n, \\ \boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left[-\sum_{i=1}^{n} m_i \pi_i^{(t)} (1 - \pi_i^{(t)}) \boldsymbol{x}_i \boldsymbol{x}_i^T \right]^{-1} \left[\sum_{i=1}^{n} (y_i - m_i \pi_i^{(t)}) \boldsymbol{x}_i \right], \\ \qquad t = 0, 1, 2, \ldots \end{cases} \tag{6.92}$$

应该用一个合理的初始估计作为 $\boldsymbol{\beta}$ 的初值 $\boldsymbol{\beta}^{(0)}$。记 $\hat{\pi}_i = y_i / m_i$, 以 $(\boldsymbol{x}_i, \hat{\pi}_i)$, $i = 1, 2, \ldots, n$ 为自变量和因变量做普通最小二乘回归得到回归系数估计可以用作 $\boldsymbol{\beta}^{(0)}$。这里不用 $\mathrm{logit}^{-1}(\hat{\pi}_i)$ 作为因变量是因为 $\hat{\pi}_i = 0$ 或 1 时 $\mathrm{logit}^{-1}(\hat{\pi}_i)$ 无定义。 □

精简最大似然估计

在最大似然估计问题中，如果能分步求得最大值，则可以减少问题的维数从而降低难度。设参数 $\boldsymbol{\theta}$ 分为 $\boldsymbol{\theta}_1$ 和 $\boldsymbol{\theta}_2$ 两部分, 如果对任意 $\boldsymbol{\theta}_1$，都能比较容易地求得

$$\arg\max_{\boldsymbol{\theta}_2} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1), \tag{6.93}$$

则

$$\max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2} l(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \max_{\boldsymbol{\theta}_1}\max_{\boldsymbol{\theta}_2} l(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \max_{\boldsymbol{\theta}_1} l(\boldsymbol{\theta}_1,\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1)), \tag{6.94}$$

问题简化为以 $\boldsymbol{\theta}_1$ 为自变量的优化问题。这样的方法称为**精简最大似然方法** (concentrated MLE)。

如果给定 $\boldsymbol{\theta}_2$ 后也很容易得到关于 $\boldsymbol{\theta}_1$ 的最大值点 $\hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2)$，那么，可以迭代地求两部分的最大值。首先取 $\boldsymbol{\theta}_1$ 的适当初值 $\boldsymbol{\theta}_1^{(0)}$，求得 $\boldsymbol{\theta}_2^{(0)} \triangleq \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^{(0)})$, 再令 $\boldsymbol{\theta}_1^{(1)} \triangleq \hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_2^{(0)})$, $\boldsymbol{\theta}_2^{(1)} \triangleq \hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_1^{(1)})$, 如此迭代直至收敛，这就构成了一种分块松弛法（见6.3.1）。这样的方法实现简单，但有可能速度比较慢。

例 6.5.2. 设总体 X 服从 $\Gamma(\alpha,\lambda)$ 分布，密度为

$$p(x;\alpha,\lambda) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x}, & x>0,\alpha>0,\lambda>0, \\ 0, & \text{其它} \end{cases} \tag{6.95}$$

设有 X 的简单随机样本 $X_1, X_2, \ldots, X_n$，则对数似然函数为

$$l(\alpha,\lambda) = n\alpha\log\lambda - n\log\Gamma(\alpha) + \alpha\sum\log X_i - \lambda\sum X_i - \sum\log X_i, \tag{6.96}$$

给定 $\alpha>0$，先关于 λ 求最大值。易见

$$\frac{\partial l}{\partial\lambda} = \frac{n\alpha}{\lambda} - \sum X_i, \tag{6.97}$$

解得稳定点 $\hat{\lambda} = \hat{\lambda}(\alpha) = \alpha/\bar{X}$, 其中 $\bar{X} = \frac{1}{n}\sum X_i$。又 $\frac{\partial^2 l}{\partial\lambda^2} = -\frac{n\alpha}{\lambda^2} < 0$, $\forall\lambda>0$, 所以 $\hat{\lambda}(\alpha)$ 是给定 α 情形下 $l(\alpha,\lambda)$ 关于自变量 λ 的最大值点。令 $\tilde{l}(\alpha) = l(\alpha,\hat{\lambda}(\alpha))$, 则

$$\begin{aligned} \tilde{l}(\alpha) =& n\alpha\log\alpha - n\log\Gamma(\alpha) + \alpha\left[\sum\log\frac{X_i}{\bar{X}} - n\right] - \sum\log X_i, \\ \tilde{l}'(\alpha) =& n\log\alpha - n\psi(\alpha) + \sum\log\frac{X_i}{\bar{X}}, \end{aligned}$$

其中 $\psi(\alpha) = \frac{d}{d\alpha}\log\Gamma(\alpha)$ 称为 digamma 函数。可以证明 $\tilde{l}'(\alpha)$ 严格单调减，方程 $\tilde{l}'(\alpha)=0$ 存在唯一解 $\hat{\alpha}$，且 $\hat{\alpha} = \arg\max_{\alpha>0}\tilde{l}(\alpha)$，从而 $(\hat{\alpha},\hat{\lambda}(\hat{\alpha}))$ 为总体参数的最大似然估计。原来二维的优化问题简化为一个一元函数 $\tilde{l}(\alpha)$ 的优化问题，可以用二分法或牛顿法求解方程 $\tilde{l}'(\alpha)=0$, 见习题25。 □

6.5.2 EM 算法

EM 算法最初用于缺失数据模型参数估计，现在已经用在许多优化问题中。设模型中包含 $\boldsymbol{X}_{\text{obs}}$ 和 $\boldsymbol{X}_{\text{mis}}$ 两个随机成分，有联合密度函数或概率函数 $f(\boldsymbol{x}_{\text{obs}},\boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta})$, $\boldsymbol{\theta}$ 为未知参数。

称 $f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta})$ 为完全数据的密度，一般具有简单的形式。实际上我们只有 $\boldsymbol{X}_{\text{obs}}$ 的观测数据 $\boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}$，$\boldsymbol{X}_{\text{mis}}$ 不能观测得到，这一部分可能是缺失观测数据，也可能是潜在影响因素。所以实际的似然函数为

$$L(\boldsymbol{\theta}) = f(\boldsymbol{x}_{\text{obs}}|\boldsymbol{\theta}) = \int f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta})\, d\boldsymbol{x}_{\text{mis}}, \tag{6.98}$$

这个似然函数通常比完全数据的似然函数复杂得多，所以很难直接从 $L(\boldsymbol{\theta})$ 求最大似然估计。

EM 算法的想法是，已经有了参数的近似估计值 $\boldsymbol{\theta}^{(t)}$ 后，假设 $(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}})$ 近似服从完全密度 $f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta}^{(t)})$, 这里 $\boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}$ 已知，所以认为 $\boldsymbol{X}_{\text{mis}}$ 近似服从由 $f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta}^{(t)})$ 导出的条件分布

$$f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)}) = \frac{f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta}^{(t)})}{f(\boldsymbol{x}_{\text{obs}}|\boldsymbol{\theta}^{(t)})}, \tag{6.99}$$

其中 $f(\boldsymbol{x}_{\text{obs}}|\boldsymbol{\theta}^{(t)})$ 是由 $f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta}^{(t)})$ 决定的边缘密度。据此近似条件分布，在完全数据对数似然函数 $\log f(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}|\boldsymbol{\theta})$ 中，把 $\boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}$ 看成已知，关于未知部分 $\boldsymbol{X}_{\text{mis}}$ 求期望，得到 $\boldsymbol{\theta}$ 的函数 $Q_t(\boldsymbol{\theta})$，再求 $Q_t(\boldsymbol{\theta})$ 的最大值点作为下一个 $\boldsymbol{\theta}^{(t+1)}$。

EM 算法每次迭代有如下的 E 步（期望步）和 M 步（最大化步）：

E 步: 计算完全数据对数似然函数的期望 $Q_t(\boldsymbol{\theta}) = E\big\{\log f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{X}_{\text{mis}}|\boldsymbol{\theta})\big\}$, 其中期望针对随机变量 $\boldsymbol{X}_{\text{mis}}$，求期望时假定 $\boldsymbol{X}_{\text{mis}}$ 服从条件密度 $f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)})$。

M 步: 求 $Q_t(\boldsymbol{\theta})$ 的最大值点，记为 $\boldsymbol{\theta}^{(t+1)}$，迭代进入下一步。

定理 6.5.1. EM 算法得到的估计序列 $\boldsymbol{\theta}^{(t)}$ 使得(6.98)中的似然函数值 $L(\boldsymbol{\theta}^{(t)})$ 单调不减。

证明 对任意参数 $\boldsymbol{\theta}$，有

$$\begin{aligned}
\ln L(\theta) =& \ln f(\boldsymbol{x}_{\text{obs}}|\theta) \cdot \int f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}} \\
=& \int \big[\log f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta}) - \log f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta})\big] f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}} \\
=& \int \log f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}}|\boldsymbol{\theta}) f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}} \\
& - \int \log f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}) f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}} \\
=& Q_t(\boldsymbol{\theta}) - \int \log f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}) f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}}
\end{aligned}$$

由信息不等式 (见习题4) 知

$$\begin{aligned}&\int \log f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{\theta}) f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}}\\ \le &\int \log f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{\theta}^{(t)}) f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}}\end{aligned}$$

又 EM 迭代使得 $Q_t(\boldsymbol{\theta}^{(t+1)}) \ge Q_t(\boldsymbol{\theta}^{(t)})$, 所以

$$\begin{aligned}\log L(\boldsymbol{\theta}^{(t+1)}) \ge & Q_t(\boldsymbol{\theta}^{(t)}) - \int \log f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{\theta}^{(t)}) f(\boldsymbol{x}_{\text{mis}}|\boldsymbol{x}_{\text{obs}},\boldsymbol{\theta}^{(t)})\, d\boldsymbol{x}_{\text{mis}}\\ = & \log L(\boldsymbol{\theta}^{(t)}).\end{aligned}$$

定理证毕。 □

在适当正则性条件下，EM 算法的迭代序列 $\boldsymbol{\theta}^{(t)}$ 依概率收敛到 $L(\boldsymbol{\theta})$ 的最大值点 $\hat{\boldsymbol{\theta}}$。

在实际问题中，往往 E 步和 M 步都比较简单，有时 E 步和 M 步都有解析表达式，这时 EM 算法实现很简单。EM 算法优点是计算稳定，可以保持原有的参数约束，缺点是收敛可能很慢，尤其是接近最大值点时可能收敛更慢。如果(6.98)中的似然函数不是凸函数，算法可能收敛不到全局最大值点，遇到这样的问题可以多取不同初值比较，用矩估计等合适的近似值作为初值。

例 6.5.3. 设总体为 N(θ,1) 分布，观测样本 $X_1,\ldots,X_n$ 有观测值 $x_1,\ldots,x_n$, 另外有 $Z_1,\ldots,Z_m$ 仅知道 $Z_j > a$, $j=1,\ldots,m$，称 Z_j 是右删失观测。假设这些观测都相互独立，则实际的似然函数为

$$L(\theta) = f(\boldsymbol{x}|\theta) = [1-\Phi(a-\theta)]^m \prod_{i=1}^{n} \phi(x_i-\theta), \tag{6.100}$$

其中 $\phi(\cdot)$ 和 $\Phi(\cdot)$ 分别为标准正态分布的分布密度和分布函数。

用 EM 算法求最大似然估计，以 $\boldsymbol{X} = (X_1,\ldots,X_n)$ 为有观测的部分，以 $\boldsymbol{Z} = (Z_1,\ldots,Z_m)$ 为没有观测的部分。完全数据的联合密度为

$$f(\boldsymbol{x},\boldsymbol{z}|\theta) = \prod_{i=1}^{n} \phi(x_i-\theta) \prod_{j=1}^{m} \phi(z_j-\theta). \tag{6.101}$$

在 E 步，设 $\theta^{(t)}$ 为参数的一个估计值，在已知 $\boldsymbol{X}=\boldsymbol{x}$ 条件下，$\boldsymbol{Z}$ 的条件密度可以从完全数据联合密度导出:

$$f(\boldsymbol{z}|\boldsymbol{x},\theta^{(t)}) = \frac{f(\boldsymbol{x},\boldsymbol{z}|\theta^{(t)})}{f(\boldsymbol{x}|\theta^{(t)})} = [1-\Phi(a-\theta^{(t)})]^{-m} \prod_{j=1}^{m} \phi(z_j-\theta^{(t)}), \tag{6.102}$$

这个条件分布可以看成是与 $\boldsymbol{X}$ 独立，且 $Z_1,\ldots,Z_m$ 独立，有共同的条件密度 $\phi(z-\theta^{(t)})/[1-\Phi(a-\theta^{(t)})]$, $z>a$。求完全数据对数似然关于 $\boldsymbol{Z}$ 的条件期望, 有

$$\begin{aligned}Q_t(\theta)=&E\log\left[\prod_{i=1}^{n}\phi(x_i-\theta)\prod_{j=1}^{m}\phi(Z_j-\theta)\right]\\=&\sum_{i=1}^{n}\log\phi(x_i-\theta)+\sum_{j=1}^{m}E\log\phi(Z_j-\theta)\end{aligned}$$

其中 Z_j 独立，有共同密度 $\phi(z-\theta^{(t)})/[1-\Phi(a-\theta^{(t)})]$,$z>a$, 所以

$$Q_t(\theta)=\sum_{i=1}^{n}\log\phi(x_i-\theta)+m\int_a^{\infty}\log\phi(z-\theta)\cdot\frac{\phi(z-\theta^{(t)})}{1-\Phi(a-\theta^{(t)})}\,dz \tag{6.103}$$

在 M 步，要求 $Q_t(\theta)$ 的最大值点。注意 $\frac{d}{dx}\log\phi(x)=\phi'(x)/\phi(x)=-x$, 得

$$\begin{aligned}Q_t'(\theta)=&\sum_{i=1}^{n}(x_i-\theta)+m\int_a^{\infty}(z-\theta)\frac{\phi(z-\theta^{(t)})}{1-\Phi(a-\theta^{(t)})}\,dz\\=&n\bar{X}-n\theta+m\int_a^{\infty}(z-\theta^{(t)})\frac{\phi(z-\theta^{(t)})}{1-\Phi(a-\theta^{(t)})}\,dz+m(\theta^{(t)}-\theta)\\=&n\bar{X}-n\theta+m\theta^{(t)}-m\theta+m\frac{1}{1-\Phi(a-\theta^{(t)})}\int_{a-\theta^{(t)}}^{\infty}u\phi(u)\,du\quad(\text{令 } u=z-\theta^{(t)})\\=&-(n+m)\theta+n\bar{X}+m\theta^{(t)}+m\frac{\phi(a-\theta^{(t)})}{1-\Phi(a-\theta^{(t)})},\end{aligned}$$

令 $Q_t'(\theta)=0$ 得

$$\theta^{(t+1)}=\frac{1}{n+m}\left[n\bar{X}+m\theta^{(t)}+m\frac{\phi(a-\theta^{(t)})}{1-\Phi(a-\theta^{(t)})}\right] \tag{6.104}$$

为 $Q_t(\theta)$ 的最大值点。选取适当初值 $\theta^{(0)}$ 按(6.104)迭代就可以求得带有右删失数据的正态总体参数 θ 的最大似然估计。 □

例 6.5.4. EM 算法可以用来估计混合分布的参数。设随机变量 $Y_1\sim \mathrm{N}(\mu_1,\delta_1)$, $Y_2\sim \mathrm{N}(\mu_2,\delta_2)$, Y_1,Y_2 独立。记 $N(\mu,\delta)$ 的密度为 $f(x|\mu,\delta)$。设随机变量 $W\sim \mathrm{b}(1,\lambda)$, $0<\lambda<1$, W 与 Y_1,Y_2 独立，令

$$X=(1-W)Y_1+WY_2, \tag{6.105}$$

则 $W=0$ 条件下 $X\sim \mathrm{N}(\mu_1,\delta_1)$, $W=1$ 条件下 $X\sim \mathrm{N}(\mu_2,\delta_2)$, 但 X 的边缘密度为

$$f(x|\boldsymbol{\theta})=(1-\lambda)f(x|\mu_1,\delta_1)+\lambda f(x|\mu_2,\delta_2), \tag{6.106}$$

其中 $\boldsymbol{\theta}=(\mu_1,\delta_1,\mu_2,\delta_2,\lambda)$。设 X 有样本 $\boldsymbol{X}=(X_1,\ldots,X_n)$, 样本值为 $\boldsymbol{x}$，实际观测数据的似然函数为

$$L(\boldsymbol{\theta})=\prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}), \tag{6.107}$$

这个函数是光滑函数但是形状很复杂，直接求极值很容易停留在局部极值点。

用 EM 算法，以 $\boldsymbol{W}=(W_1,\ldots,W_n)$ 为没有观测到的部分, 完全数据的似然函数和对数似然函数为

$$\begin{aligned}\tilde{L}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{W})=&\prod_{W_i=0} f(x_i|\mu_1,\delta_1)\prod_{W_i=1} f(x_i|\mu_2,\delta_2)\ \lambda^{\sum_{i=1}^n W_i}(1-\lambda)^{n-\sum_{i=1}^n W_i},\\ \tilde{l}(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{W})=&\sum_{i=1}^{n}\Big[(1-W_i)\log f(x_i|\mu_1,\delta_1)+W_i\log f(x_i|\mu_2,\delta_2)\Big]\\ &+\left(\sum_{i=1}^{n}W_i\right)\log\lambda+\left(n-\sum_{i=1}^{n}W_i\right)\log(1-\lambda).\end{aligned} \tag{6.108}$$

在 E 步，设已有 $\boldsymbol{\theta}$ 的近似值 $\boldsymbol{\theta}^{(t)}=(\mu_1^{(t)},\delta_1^{(t)},\mu_2^{(t)},\delta_2^{(t)},\lambda^{(t)})$, 以 $\boldsymbol{\theta}^{(t)}$ 为分布参数，在 $\boldsymbol{X}=\boldsymbol{x}$ 条件下，W_i 的条件分布为

$$\begin{aligned}\gamma_i^{(t)}\triangleq&P(W_i=1|\boldsymbol{x},\boldsymbol{\theta}^{(t)})=P(W_i=1|X_i=x_i,\boldsymbol{\theta}^{(t)})\\ =&\frac{\lambda^{(t)}f(x_i|\mu_2^{(t)},\delta_2^{(t)})}{(1-\lambda^{(t)})f(x_i|\mu_1^{(t)},\delta_1^{(t)})+\lambda^{(t)}f(x_i|\mu_2^{(t)},\delta_2^{(t)})}.\end{aligned} \tag{6.109}$$

这里的推导类似于逆概率公式。利用 W_i 的条件分布求完全数据对数似然的期望，得

$$\begin{aligned}Q_t(\boldsymbol{\theta})=&\sum_{i=1}^{n}\Big[(1-\gamma_i^{(t)})\log f(x_i|\mu_1,\delta_1)+\gamma_i^{(t)}\log f(x_i|\mu_2,\delta_2)\Big]\\ &+\left(\sum_{i=1}^{n}\gamma_i^{(t)}\right)\log\lambda+\left(n-\sum_{i=1}^{n}\gamma_i^{(t)}\right)\log(1-\lambda).\end{aligned} \tag{6.110}$$

令 $\nabla Q_t(\boldsymbol{\theta})=\mathbf{0}$，求得 $Q_t(\boldsymbol{\theta})$ 的最大值点 $\boldsymbol{\theta}^{(t+1)}$ 为

$$\begin{cases}\mu_1^{(t+1)}=\frac{\sum_{i=1}^n(1-\gamma_i^{(t)})x_i}{\sum_{i=1}^n(1-\gamma_i^{(t)})}\\ \delta_1^{(t+1)}=\frac{\sum_{i=1}^n(1-\gamma_i^{(t)})(x_i-\mu_1^{(t+1)})^2}{\sum_{i=1}^n(1-\gamma_i^{(t)})}\\ \mu_2^{(t+1)}=\frac{\sum_{i=1}^n\gamma_i^{(t)}x_i}{\sum_{i=1}^n\gamma_i^{(t)}}\\ \delta_2^{(t+1)}=\frac{\sum_{i=1}^n\gamma_i^{(t)}(x_i-\mu_2^{(t+1)})^2}{\sum_{i=1}^n\gamma_i^{(t)}}\\ \lambda^{(t+1)}=\frac{1}{n}\sum_{i=1}^n\gamma_i^{(t)}\end{cases} \tag{6.111}$$

适当选取初值 $\boldsymbol{\theta}^{(0)}$ 用(6.109)和(6.111)迭代就可以计算 $\boldsymbol{\theta}$ 的最大似然估计。 □

6.5.3 非线性回归

首先要注意有些非线性回归可以轻易地转换为线性回归。比如，$y = Ae^{-\beta x}$ 可以变成 $\log y = a - \beta x$, 其中 $a = \log A$。又如，$y = A\cos(2\pi ft + \varphi)(f$ 已知) 可以变成 $y = a\cos(2\pi ft) + b\sin(2\pi ft)$, 其中 $a = A\cos\varphi$, $b = -A\sin\varphi$。

考虑如下的非线性回归问题

$$Y_i = \varphi(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i,\ i = 1, 2, \ldots, n,$$

其中 ε_i 独立同 $\mathrm{N}(0,\sigma^2)$ 分布，$\boldsymbol{\beta} \in \mathbb{R}^p$ 为未知参数向量，Y_i 为因变量，$\boldsymbol{x}_i$ 为非随机的自变量，φ 为关于 $\boldsymbol{\beta}$ 非线性的函数。

记 $\gamma = \sigma^{-2}$(这可以使得似然函数偏导数形式更简单), $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $g_i(\boldsymbol{\beta}) = \varphi(\boldsymbol{x}_i, \boldsymbol{\beta})$, $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{\beta}) = (g_1(\boldsymbol{\beta}), \ldots, g_n(\boldsymbol{\beta}))^T$,

$$G = G(\boldsymbol{\beta}) = \left(\frac{\partial g_i(\boldsymbol{\beta})}{\partial \beta_j}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}},$$

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \big[Y_i - g_i(\boldsymbol{\beta})\big]^2 = \|\boldsymbol{Y} - \boldsymbol{g}(\boldsymbol{\beta}\|^2.$$

最小化 $S(\boldsymbol{\beta})$ 可以得到参数 $\boldsymbol{\beta}$ 的最小二乘估计，容易验证最小二乘估计也是最大似然估计。求非线性最小二乘估计可以用通常的牛顿法、拟牛顿法、Nelder-Mead 等方法，下面结合非线性回归模型最大似然估计的特点讨论更为高效的参数估计算法。

上述模型的对数似然函数为 (省略常数项)

$$l(\boldsymbol{\beta}, \gamma) = \frac{n}{2}\log\gamma - \frac{\gamma}{2}S(\boldsymbol{\beta}),$$

梯度和海色阵为

$$\nabla l(\boldsymbol{\beta}, \gamma) = \begin{pmatrix} \gamma G^T(\boldsymbol{Y} - \boldsymbol{g}) \\ \frac{n}{2\gamma} - \frac{1}{2}S(\boldsymbol{\beta}) \end{pmatrix},$$

$$\nabla^2 l(\boldsymbol{\beta}, \gamma) = \begin{pmatrix} \gamma\sum_{i=1}^{n}[Y_i - g_i(\boldsymbol{\beta})]\nabla^2 g_i(\boldsymbol{\beta}) - \gamma G^T G & G^T(\boldsymbol{Y} - \boldsymbol{g}) \\ (\boldsymbol{Y} - \boldsymbol{g})^T G & -\frac{n}{2}\gamma^{-2} \end{pmatrix},$$

由此得 $\boldsymbol{Y}$ 的信息阵为

$$I_n(\boldsymbol{\beta}, \gamma) = E\big[-\nabla^2 l(\boldsymbol{\beta}, \gamma)\big] = \begin{pmatrix} \gamma G^T G & 0 \\ 0 & \frac{n}{2}\gamma^{-2} \end{pmatrix},$$

可见信息阵比海色阵简单得多，用得分法迭代估计参数比牛顿法更简单，迭代公式为

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left[G^T(\boldsymbol{\beta}^{(t)})G(\boldsymbol{\beta}^{(t)})\right]^{-1} G^T(\boldsymbol{\beta}^{(t)}) \left[\boldsymbol{Y} - \boldsymbol{g}(\boldsymbol{\beta}^{(t)})\right], \tag{6.112}$$

这种方法称为高斯–牛顿法。如果拟合误差 $Y_i - g_i(\boldsymbol{\beta})$ 都比较小，这时 $-I_n(\boldsymbol{\beta}, \gamma)$ 与海色阵 $\nabla^2 l(\boldsymbol{\beta}, \gamma)$ 会很接近，这种情况下高斯–牛顿法具有很好的效果。与牛顿法的缺点类似，高斯–牛顿法有可能步长过大，使得 $S(\boldsymbol{\beta})$ 变大而不是变小，这时可以在(6.112)中增加一个步长 α，从 $\alpha = 1$ 开始每次步长减半直到迭代使 $S(\boldsymbol{\beta})$ 变小。

当拟合误差 $Y_i - g_i(\boldsymbol{\beta})$ 较大时，高斯–牛顿法效果会变差。把公式(6.112)改为

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left[G^T(\boldsymbol{\beta}^{(t)})G(\boldsymbol{\beta}^{(t)}) + \lambda_t I_p\right]^{-1} G^T(\boldsymbol{\beta}^{(t)}) \left[\boldsymbol{Y} - \boldsymbol{g}(\boldsymbol{\beta}^{(t)})\right], \tag{6.113}$$

其中 $\lambda_t \geq 0$, $\lambda_t \geq 0$ 可以迭代地更新。此算法称为 LMF(Levenberg-Marquardt-Fletcher) 算法。公式(6.113) 可以看成是取搜索方向为高斯–牛顿法的搜索方向和最速下降法的搜索方向之间的一个方向, 记搜索方向为

$$\boldsymbol{d}^{(t)} = \left[G^T(\boldsymbol{\beta}^{(t)})G(\boldsymbol{\beta}^{(t)}) + \lambda_t I_p\right]^{-1} G^T(\boldsymbol{\beta}^{(t)}) \left[\boldsymbol{Y} - \boldsymbol{g}(\boldsymbol{\beta}^{(t)})\right] \tag{6.114}$$

并记

$$\begin{aligned}
\Delta S_t &= S(\boldsymbol{\beta}^{(t)}) - S(\boldsymbol{\beta}^{(t)} + \boldsymbol{d}^{(t)}), \\
\Delta q_t &= (\boldsymbol{d}^{(t)})^T \left[\lambda_t \boldsymbol{d}^{(t)} + G^T(\boldsymbol{\beta}^{(t)}) \left(\boldsymbol{Y} - \boldsymbol{g}(\boldsymbol{\beta}^{(t)})\right)\right],
\end{aligned}$$

LMF 算法如下:

取初值 $\boldsymbol{\beta}^{(0)}$, $\lambda_0 > 0$, 精度 $\epsilon > 0$, $t \leftarrow 0$
until (迭代收敛) {
  用(6.114)求 $\boldsymbol{d}^{(t)}$
  计算 $\gamma_t = \Delta S_t / \Delta q_t$
  if($\gamma_t < 0.25$) {
    $\lambda_{t+1} \leftarrow 4\lambda_t$
  } **else if**($\lambda_t > 0.75$) {
    $\lambda_{t+1} \leftarrow \frac{1}{2}\lambda_t$
  }
  if($\gamma_t \leq 0$) {
    $\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)}$
  } **else** {

$\boldsymbol{\beta}^{(t+1)} \leftarrow \boldsymbol{\beta}^{(t)} + \boldsymbol{d}^{(t)}$

}

$t \leftarrow t+1$

} # until

输出 $\boldsymbol{\beta}^{(t)}$ 作为最小二乘估计

参见高立 (2014)[3] §5.3。

习题六

1. 设 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$，证明如下的 Cauchy-Schwarz 不等式:

$$\big|(\boldsymbol{x}, \boldsymbol{y})\big| \leq \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|$$

其中 $(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n x_i y_i$, $\|\boldsymbol{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$，不等式中等号成立当且仅当存在实数 a, b 使得 $a\boldsymbol{x} + b\boldsymbol{y} = 0$。

2. 证明凸集性质 (1)–(10)。

3. 证明凸规划问题的局部最优解一定是全局最优解。

4. 设 $f(\boldsymbol{x})$, $g(\boldsymbol{x})$ 是定义在集合 A 上的两个密度，$f(\boldsymbol{x})$, $g(\boldsymbol{x})$ 在 A 上都为正值。证明如下信息不等式:

$$\int_A [\log f(\boldsymbol{x})] f(\boldsymbol{x})\, d\boldsymbol{x} \geq \int_A [\log g(\boldsymbol{x})] f(\boldsymbol{x})\, d\boldsymbol{x}.$$

5. 对例6.1.11, 证明当 X 不满秩时方程 $X^T X\boldsymbol{\beta} = X^T Y$ 有无穷多解, 任一个解 $\boldsymbol{\beta}^*$ 都是 $Q(\boldsymbol{\beta})$ 的全局最小值点。

6. 在定理6.1.6条件下，如果 $\{\nabla c_i(\boldsymbol{x}^*), i = 1, \ldots, p+r\}$ 构成线性无关向量组，则 $\boldsymbol{\lambda}^*$ 唯一。

7. 对约束优化问题

$$\begin{cases} \arg\min f(x_1, x_2) \triangleq 4x_1 - 3x_2, \quad \text{s.t.} \\ 4 - x_1 - x_2 \geq 0, \\ x_2 + 7 \geq 0, \\ -(x_1 - 3)^2 + x_2 + 1 \geq 0 \end{cases}$$

求其 KKT 点，并根据二阶条件判断 KKT 点是否最小值点。

8. 设一元函数 $f(x)=\frac{4}{3}\log(1+x)-x$ 定义在 $(0,1)$ 中，先求其最大值点的精确值，然后编写 R 程序，分别用 0.618 法和二分法求其最大值点, 比较 5 次和 10 次迭代后，两种方法的绝对误差大小。

9. 对例6.2.1，编写 R 函数，给定 $n,\bar{X},S,U,1-\alpha$ 后计算 K 的 $1-\alpha$ 置信下限。

10. 对二次多项式 $\varphi(x)=\beta_0+\beta_1x+\beta_2x^2$，$\beta_2>0$, β_0,β_1,β_2 都未知，若已知 $a<c<b$ 和 $y_a=\varphi(a)$, $y_c=\varphi(c)$, $y_b=\varphi(b)$, 求 $\varphi(\cdot)$ 的最小值点。

11. 对二次多项式函数 $\varphi(x)=\beta_0+\beta_1x+\beta_2x^2$，设 $\beta_2>0$, β_0,β_1,β_2 未知。若已知 $x_1\neq x_2$ 和 $y_1=\varphi(x_1)$, $y_2=\varphi(x_2)$，$y_1'=\varphi'(x_1)$，求 $\varphi(x)$ 的最小值点。

12. 对二次多项式函数 $\varphi(x)=\beta_0+\beta_1x+\beta_2x^2$，设 $\beta_2>0$, β_0,β_1,β_2 未知。若已知 $x_1\neq x_2$ 和 $y_1=\varphi(x_1)$, $y_1'=\varphi'(x_1)$，$y_2'=\varphi'(x_2)$，求 $\varphi(x)$ 的最小值点。

13. 设函数 $f(\boldsymbol{x})$ 定义在 $\mathbb{R}^d$ 上，有一阶偏导数，$\boldsymbol{u}\in\mathbb{R}^d$, $\|\boldsymbol{u}\|=1$。令 $h(\alpha)=f(\boldsymbol{x}+\alpha\boldsymbol{u})$, $\alpha\geq 0$, 证明 $h'(0)=\nabla f(\boldsymbol{x})^T\boldsymbol{u}$。

14. 用 R 语言编写牛顿法的程序，当没有输入偏导数函数时，程序用数值微分方法近似计算偏导数。

15. 用 R 语言编写 BFGS 方法的程序，当没有输入偏导数函数时，程序用数值微分方法近似计算偏导数。

16. 用 R 语言编写 Nelder-Mead 算法的程序。

17. 用外点罚函数法求解如下约束优化问题:

$$\begin{cases}\arg\min x_1^2+x_2^2, & \text{s.t.}\\ x_1-x_2+1=0.\end{cases}$$

18. 用外点罚函数法求解如下约束优化问题:

$$\begin{cases}\arg\min x_1^2+x_2^2, & \text{s.t.}\\ -x_1+x_2-1\geq 0.\end{cases}$$

19. 用内点罚函数法求解上一题目。

20. 设总体 X 服从如下的 logistic 分布:

$$f(x|\theta) = \frac{\exp(-(x-\theta))}{(1+\exp(-(x-\theta)))^2},\ -\infty < x < \infty,\ -\infty < \theta < \infty,$$

设 $\boldsymbol{X} = (X_1, \ldots, X_n)$ 为 X 的简单随机样本。

(1) 写出对数似然函数 $l(\theta)$ 和导数 $l'(\theta)$。

(2) 证明方程 $l'(\theta) = 0$ 存在唯一解，解为最大似然估计。

(3) 写出用牛顿法求解方程 $l'(\theta) = 0$ 的迭代公式。

(4) 设一个样本为 $-0.63, 0.18, -0.84, 1.6, 0.33, -0.82, 0.49, 0.74, 0.58, -0.31$, 计算 $\hat{\theta}$。

21. 设总体 X 取值于正整数集合 $\mathbb{Z}_+ \triangleq \{1, 2, 3, \ldots\}$，有概率函数

$$p(x;\theta) = \theta^x x^{-1}[-\log(1-\theta)]^{-1},\ x = 1, 2, \ldots,$$

其中 $\theta \in (0, 1)$ 为未知参数。设 $x_1, x_2, \ldots, x_n$ 为 X 的简单随机样本，设 $\sum_{i=1}^n x_n = s$。

(1) 令 $f(\theta)$ 为负对数似然函数，写出 $f(\theta)$ 和 $g(\theta) = f'(\theta)$ 以及 $g'(\theta)$ 的表达式。

(2) 设 $n = 10$, $s = 15$, 编写 R 程序，作 $f(x)$ 和 $g(x)$ 的图像;

(3) 编写 R 程序，分别用二分法、牛顿法和割线法求 θ 的最大似然估计 (即 $f(\theta)$ 的最小值点)。

22. 考虑如下最大似然估计计算问题。设 X 取值于 $\{1, 2, 3, 4\}$, 分布概率为 $p(x|\theta_1, \theta_2) \triangleq \pi_x(\theta_1, \theta_2)$:

$$\begin{aligned} \pi_1(\theta_1, \theta_2) &= 2\theta_1\theta_2, & \pi_2(\theta_1, \theta_2) &= \theta_1(2 - \theta_1 - 2\theta_2), \\ \pi_3(\theta_1, \theta_2) &= \theta_2(2 - \theta_2 - 2\theta_1), & \pi_4(\theta_1, \theta_2) &= (1 - \theta_1 - \theta_2)^2. \end{aligned}$$

设 n 次试验得到的 X 值有 n_j 个 $j (j = 1, 2, 3, 4)$，则对数似然函数为 (去掉了与参数无关的加性常数)

$$l(\theta_1, \theta_2) = \sum_{j=1}^4 n_j \log \pi_j(\theta_1, \theta_2)$$

设 $n = 435$, $n_1 = 17$, $n_2 = 182$, $n_3 = 60$, $n_4 = 176$。编写 R 程序，分别用牛顿法、阻尼牛顿法、BFGS 方法、得分法和 Nelder-Mead 方法求最大似然估计，比较这几种方法的收敛速度。

23. 设 $Y_i \sim \mathrm{B}(m_i, \mathrm{logit}^{-1}(\beta_0+\beta_1 x_i))$, $i=1,\ldots,n$ 相互独立，$n=4$，(m_i, y_i, x_i) 的 4 组数据为 $(55,0,7)$, $(157,2,14)$, $(159,7,27)$, $(16,3,51)$。编写 R 程序用牛顿法求 $\boldsymbol{\beta}=(\beta_0,\beta_1)$ 的最大似然估计。写出模型的对数似然函数及其梯度、海色阵，写出最大似然估计的牛顿法迭代公式。

24. 设随机变量 $Y_i \sim \mathrm{Poisson}(\exp[\boldsymbol{\beta}^T \boldsymbol{x}_i])$, $i=1,2,\ldots,n$, $Y_1,\ldots,Y_n$ 相互独立，$\boldsymbol{\beta}$ 为未知参数向量，自变量 $\boldsymbol{x}_i, i=1,\ldots,n$ 已知。写出模型的对数似然函数及其梯度、海色阵，写出最大似然估计的牛顿法迭代公式。

25. 写出例6.5.2中的简化似然函数 $\tilde{l}(\alpha)$，编写 R 程序，分别用二分法和牛顿法求 $\tilde{l}(\alpha)$ 的最大值点。设有样本 $0.08, 0.36, 0.35, 0.21, 0.39, 0.25, 0.23, 0.11, 0.07, 0.08$, 求参数最大似然估计。

26. 设总体 X 服从威布尔分布 $\mathrm{W}(\alpha,\eta)$，密度函数为

$$p(x;\alpha,\eta)=\begin{cases}\frac{\alpha}{\eta}\left(\frac{x}{\eta}\right)^{\alpha-1}\exp\left\{-\left(\frac{x}{\eta}\right)^{\alpha}\right\}, & x>0,\alpha>0,\eta>0,\\ 0, & \text{其它}\end{cases}$$

设 $X_1, X_2, \ldots, X_n$ 为 X 的简单随机样本。

(1) 求对数似然函数 $l(\alpha,\eta)$ 及其梯度 $\nabla l(\alpha,\eta)$、海色阵 $\nabla^2 l(\alpha,\eta)$;

(2) 编写 R 程序，用牛顿法求最大似然估计；

(3) 用分步求最大值的方法，先对固定 α 求最大值点 $\hat{\eta}(\alpha)$，再关于 α 求最大值，编写 R 程序实现算法；

(4) 用随机模拟方法生成多组样本，比较两种算法方法的收敛速度。比较随机模拟得到的最大似然估计的方差与牛顿法结束迭代时用海色阵得到的渐近方差。

27. 设 X 取值于 $\{1,2,3,4\}$, 分布概率为 $p(x|\theta) \triangleq \pi_x(\theta)$:

$$\pi_1(\theta)=\frac{1}{4}(2+\theta), \qquad \pi_2(\theta)=\frac{1}{4}(1-\theta),$$
$$\pi_3(\theta)=\frac{1}{4}(1-\theta), \qquad \pi_4(\theta)=\frac{1}{4}\theta.$$

设 n 次试验得到的 X 值有 n_j 个 $j(j=1,2,3,4)$，则对数似然函数为 (去掉了与参数无关的加性常数)

$$l(\theta)=n_1\log(2+\theta)+(n_2+n_3)\log(1-\theta)+n_4\log\theta.$$

(1) 令 $l'(\theta)=0$ 得到一个关于 θ 的二次方程，由此写出 $\arg\max l(\theta)$ 的解析表达式。

(2) 设结果中取 1 的进一步分为 11 和 12 两个部分，取 11 的概率为 $\frac{1}{2}$，取 12 的概率为 $\frac{1}{4}\theta$，设 (Z_1,Z_2) 为没有观测到的取 11 和 12 的次数。写出这时包括 (Z_1,Z_2,n_2,n_3,n_4) 的完全数据似然函数，用 EM 算法求 $l(\theta)$ 的最大值点。

(3) 设 $n=400$, $n_1=480$, $n_2=120$, $n_3=110$, $n_4=90$。编写 R 程序用 EM 算法计算最大似然估计，与解析表达式的计算结果进行比较。

28. 考虑如下非线性回归模型：

$$y_i=A\cos(\omega x_i+\theta)+\varepsilon_i,\ i=1,2,\ldots,n,$$

其中 ε_i 独立同 $\mathrm{N}(0,\sigma^2)$ 分布，A,ω,θ,σ^2 为未知参数。设 $n=100$, $x_i=2\pi i/n$, $A=10$, $\omega=2$, $\theta=0.5$, $\sigma^2=2^2$。编写 R 程序，生成随机模拟样本, 分别用高斯–牛顿法和 LMF 方法给出参数最小二乘估计。

29. 考虑删失数据的参数估计问题。设总体 Y 服从指数分布 $\mathrm{Exp}(\theta^{-1})$, $\theta=EY$。设 $u_1,u_2,\ldots,u_n$ 是 Y 的一组简单随机样本，$v_1,v_2,\ldots,v_m$ 是 Y 的与前一组样本独立的另一组简单随机样本，但是 $v_1,v_2,\ldots,v_m$ 的具体值未知，只知道 $v_1,v_2,\ldots,v_m$ 中有 r 个 $(0\le r\le m)$ 个在时刻 s 之前失效了，其余 $m-r$ 个到时刻 s 都没有失效。推导用 EM 算法计算 θ 的最大似然估计的算法。

参考文献

[1] 陈家鼎、孙山泽、李东风、刘力平 (2006) 数理统计学讲义. 高等教育出版社.

[2] 高惠璇 (1995) 统计计算. 北京大学出版社.

[3] 高立 (2014) 数值最优化方法. 北京大学出版社.

[4] 关治, 陆金甫 (1998) 数值分析基础. 高等教育出版社.

[5] 何书元 (2003) 应用时间序列分析. 北京大学出版社

[6] Hoeting, J.A. and Givens, G.H.(2009) 计算统计. 人民邮电出版社.

[7] 茆诗松, 王静龙, 濮晓龙 (2006) 高等数理统计 第二版. 高等教育出版社.

[8] Ross, S.M.(2007) 统计模拟, 第四版, 人民邮电出版社

[9] 肖筱南 (2003) 现代数值计算方法. 北京大学出版社.

[10] 徐成贤、陈志平、李乃成 (2002) 近代优化方法. 科学出版社.

[11] 徐仲、张凯院、陆全、冷国伟 (2005) 矩阵论简明教程. 第二版. 科学出版社.

[12] Anderson, E. et al(1999) *LAPACK Users' Guide*. Third Edition. SIAM. 有网上版本 http://www.netlib.org/lapack/lug/lapack_lug.html.

[13] Becker, R.A. and Chambers, J.M.(1984), *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth Advanced Books Program, Belmont CA

[14] Becker, R.A., Chambers, J.M. and Wilks, A.R. (1988), *The New S Language*. Chapman and Hall, New York.

[15] Chambers, J.M. and Hastie, T. (1992), *Statistical Models in S*. Chapman and Hall, New York.

[16] Cowles, M. K.(2013) *Applied Bayesian Statistics With R and OpenBUGS Examples*. Springer

[17] Fang, K.T. and Wang, Y.(1994) *Number-Theoretic Methods in Statistics*. Chapman & Hall

[18] Gentle, J.E.(2002) *Elements of Computational Statistics*. Springer Science + Business Media, Inc.

[19] Gentle, J.E.(2007) *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer.

[20] Gentle, J.E.(2009) *Computational Statistics*. Springer, 2009.

[21] Hogg, R.V., McKean, J.W., Craig, A.T.(2012) *Introduction to Mathematical Statistics*, 7th Ed, China Machine Press

[22] Hyndman, R.J. and Fan, Y.(1996) Sample quantiles in statistical packages. *American Statistician*, 50:361–365.

[23] IMSL Inc.(1985) IMSL Quality Mathematical And Statistical FORTRAN Subroutines for IBM Personal Computers. *IEEE Micro*, V. 5, NO. 2, p. 97.

[24] Kitagawa, G.(1996) Monte Carlo filter and smoother for non-gaussian nonlinear state-space model. *J. of Computational and Graphical Statistics*, V. 5, 1-25

[25] Lange, K.(2013) *Optimization*. Springer

[26] Lemieux, C.(2009) *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer

[27] Liu, J.S., Chen, R. and Wong, W. H.(1998) Rejection Control and Sequential Importance Sampling. *JASA*, V. 93, 1022-1031

[28] Liu, J.S.(2001) *Monte Carlo Strategies in Scientific Computing*. Springer

[29] Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009) The BUGS project: Evolution, critique and future directions (with discussion), *Statistics in Medicine* V. 28, 3049–3082.

[30] Lunn, D., Thomas, A., Best, N., Spiegelhalter, D. (2000). WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, V. 10, 325–337

[31] Monahan, J.F.(2001) *Numerical Methods of Statistics*. Cambridge University Press.

[32] Nelder, J.A. and Mead, R. (1965) A simplex method for function minimization, *Comput. J.*, 7, pp. 308–313.

[33] Ntzoufras, I.(2009) *Bayesian Modeling Using WinBUGS.* Wiley

[34] Ross, S.M.(2013) *Simulation* 5th Ed, Elsevier Inc.

[35] SAS Institute Inc(2010) *Base SAS 9.2 Procedures Guide: Statistical Procedures* 3rd ed. Cary, NC: SAS Institute Inc.

[36] Stewart, G.W.(1973) *Introduction to Matrix Computations*. New York: Academic Press.

[37] Sturges, H.A.(1926) The choice of a class interval. *Journal of the American Statistical Association*, 21:65–66, 1926.