

Preface

前言

感谢

首先感谢大家的信任。

作者仅仅是在学习应用数学科学和机器学习算法时，多读了几本数学书，多做了些思考和知识整理而已。知者不言，言者不知。知者不博，博者不知。水平有限，把自己有限所学所思斗胆和大家分享，作者权当无知者无畏。希望大家在 B 站视频下方和 Github 多提意见，让这套书成为作者和读者共同参与创作的优质作品。

特别感谢清华大学出版社的栾大成老师。从选题策划、内容创作、装帧设计，栾老师事无巨细、一路陪伴。每次和栾老师交流，我都能感受到他对优质作品的追求、对知识分享的热情。

出来混总是要还的

曾几何时，考试是我们学习数学的唯一动力。考试是头悬梁的绳，是锥刺股的锥。我们中的绝大多数人从小到大为各种考试埋头题海，数学味同嚼蜡，甚至让人恨之入骨。

数学给我们带来了无尽的折磨。我们憎恨数学，恐惧数学，恨不得一走出校门就把数学抛之脑后、老死不相往来。

可悲可笑的是，我们其中很多人可能会在毕业的五年或十年以后，因为工作需要，不得不重新学习微积分、线性代数、概率统计，悔恨当初没有学好数学、走了很多弯路、没能学以致用，从而迁怒于教材和老师。

这一切不能都怪数学，值得反思的是我们学习数学的方法、目的。

再给自己一个学数学的理由

为考试而学数学，是被逼无奈的举动。而为数学而数学，则又太过高尚而遥不可及。

相信对于绝大部分的我们来说，数学是工具、是谋生手段，而不是目的。我们主动学数学，是想用数学工具解决具体问题。

现在，这套书给大家一个“学数学、用数学”的全新动力——数据科学、机器学习。

数据科学和机器学习已经深度融合到我们生活的方方面面，而数学正是开启未来大门的钥匙。不是所有人生来都握有一副好牌，但是掌握“数学 + 编程 + 机器学习”绝对是王牌。这次，学习数学不再是为了考试、分数、升学，而是投资时间、自我实现、面向未来。

未来已来，你来不来？

本套丛书如何帮到你

为了让大家学数学、用数学，甚至爱上数学，作者可谓颇费心机。在创作这套书时，作者尽量克服传统数学教材的各种弊端，让大家学习时有兴趣、看得懂、有思考、更自信、用得着。

为此，丛书在内容创作上突出以下几个特点：

- ◀ **数学 + 艺术**——全彩图解，极致可视化，让数学思想跃然纸上、生动有趣、一看就懂，同时提高大家的数据思维、几何想象力、艺术感；
- ◀ **零基础**——从零开始学习 Python 编程，从写第一行代码到搭建数据科学和机器学习应用；
- ◀ **知识网络**——打破数学板块之间的壁垒，让大家看到数学代数、几何、线性代数、微积分、概率统计等板块之间的联系，编织一张绵密的数学知识网络；
- ◀ **动手**——授人以鱼不如授人以渔，和大家一起写代码、用 Streamlit 创作数学动画、交互 App；
- ◀ **学习生态**——构造自主探究式学习生态环境“微课视频 + 纸质图书 + 电子图书 + 代码文件 + 可视化工具 + 思维导图”，提供各种优质学习资源；
- ◀ **理论 + 实践**——从加减乘除到机器学习，丛书内容安排由浅入深、螺旋上升，兼顾理论和实践；在编程中学习数学，学习数学时解决实际问题。

虽然本书标榜“从加减乘除到机器学习”，但是建议读者朋友们至少具备高中数学知识。如果读者正在学习或曾经学过大学数学（微积分、线性代数、概率统计），这套书就更容易读了。

聊聊数学

数学是工具。锤子是工具，剪刀是工具，数学也是工具。

数学是思想。数学是人类思想的高度抽象的结晶体。在其冷酷的外表之下，数学的内核实际上就是人类朴素的思想。学习数学时，知其然，更要知其所以然。不要死记硬背公式定理，理解背后的数学思想才是关键。如果你能画一幅图、用大白话描述清楚一个公式、一则定理，这就说明你真正理解了它。

数学是语言。就好比世界各地不同种族有自己的语言，数学则是人类共同的语言和逻辑。数学这门语言极其精准、高度抽象，放之四海而皆准。虽然我们中绝大多数人没有被数学女神选中，不能为人类的对数学认知开疆扩土；但是，这丝毫不妨碍我们使用数学这门语言。就好比，我们不会成为语言学家，我们完全可以使用母语和外语交流。

数学是体系。代数、几何、线性代数、微积分、概率统计、优化方法等等，看似一个个孤岛，实际上都是数学网络的一条条织线。建议大家学习时，特别关注不同数学板块之间的联系，见树，更要见林。

数学是基石。拿破仑曾说“数学的日臻完善和这个国强民富息息相关。”数学是科学进步的根基，是经济繁荣的支柱，是保家卫国的武器，是探索星辰大海的航船。

数学是艺术。数学和音乐、绘画、建筑一样，都是人类艺术体验。通过可视化工具，我们会在看似枯燥的公式、定理、数据背后，发现数学之美。

数学是历史，是人类共同记忆体。“历史是过去，又属于现在，同时在指引未来。”数学是人类的集体学习思考，她把人的思维符号化、形式化，进而记录、积累、传播、创新、发展。从甲

骨、泥板、石板、竹简、木牍、纸草、羊皮卷、活字印刷、纸质书，到数字媒介，这一过程持续了数千年，至今绵延不息。

数学是无穷无尽的**想象力**，是人类的**好奇心**，是自我挑战的**毅力**，是一个接着一个的**问题**，是看似荒诞不经的**猜想**，是一次次胆大包天的**批判性思考**，是敢于站在前人的肩膀之上的**勇气**，是孜孜不倦地延展人类认知边界的**不懈努力**。

家园、诗、远方

诺瓦利斯曾说：“哲学就是怀着一种乡愁的冲动到处去寻找家园。”

在纷繁复杂的尘世，数学纯粹的就像精神的世外桃源。数学是，一束光，一条巷，一团不灭的希望，一股磅礴的力量，一个值得寄托的避风港。

打破陈腐的锁链，把功利心暂放一边，我们一道怀揣一分乡愁、心存些许诗意、踩着艺术维度，投入数学张开的臂膀，驶入她色彩斑斓、变幻无穷的深港，感受久违的归属，一睹更美、更好的远方。

Acknowledgement

致谢

To my parents.

谨以此书献给我的母亲父亲

How to Use the Book

使用本书

丛书资源

本系列丛书提供的配套资源有以下几个：

- ◀ 纸质图书；
- ◀ PDF 文件，方便移动终端学习；请大家注意，纸质图书经过出版社五审五校修改，内容细节上会和 PDF 文件有出入。
- ◀ 每章提供思维导图，纸质书提供全书思维导图海报；
- ◀ Python 代码文件，直接下载运行，或者复制、粘贴到 Jupyter 运行；
- ◀ Python 代码中有专门用 Streamlit 开发数学动画和交互 App 的文件；
- ◀ 微课视频，强调重点、讲解难点、聊聊天。

在纸质书中为了方便大家查找不同配套资源，作者特别设计了如下几个标识。



微课视频

本书配套微课视频均发布在 B 站——生姜 DrGinger：

- ◀ <https://space.bilibili.com/513194466>

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

微课视频是以“聊天”的方式，和大家探讨某个数学话题的重点内容，讲讲代码中可能遇到的难点，甚至侃侃历史、说说时事、聊聊生活。

本书配套的微课视频目的是引导大家自主编程实践、探究式学习，并不是“照本宣科”。

纸质图书上已经写得很清楚的内容，视频课程只会强调重点。需要说明的是，图书内容不是视频的“逐字稿”。

代码文件

本系列丛书的 Python 代码文件下载地址为：

► <https://github.com/Visualize-ML>

Python 代码文件会不定期修改，请大家注意更新。图书配套的 PDF 文件和勘误也会上传到这个 GitHub 账户。因此，建议大家注册 GitHub 账户，给书稿文件夹标星 (star) 或分支克隆 (fork)。

考虑再三，作者还是决定不把代码全文印在纸质书中，以便减少篇幅，节约用纸。

本书编程实践例子中主要使用“鸢尾花数据集”，数据来源是 Scikit-learn 库、Seaborn 库。此外，系列丛书封面设计致敬梵高《鸢尾花》，要是给本系列丛书起个昵称的话，作者乐见“鸢尾花书”。

App 开发

本书几乎每一章都至少有一个用 Streamlit 开发的 App，用来展示数学动画、数据分析、机器学习算法。

Streamlit 是个开源的 Python 库，能够方便快捷搭建、部署交互型网页 App。Streamlit 非常简单易用、很受欢迎。Streamlit 兼容目前主流的 Python 数据分析库，比如 NumPy、Pandas、Scikit-learn、PyTorch、TensorFlow 等等。Streamlit 还支持 Plotly、Bokeh、Altair 等交互可视化库。

本书中很多 App 设计都采用 Streamlit + Plotly 方案。此外，本书专门配套教学视频手把手和大家一起做 App。

大家可以参考如下页面，更多了解 Streamlit：

► <https://streamlit.io/gallery>

► <https://docs.streamlit.io/library/api-reference>

实践平台

本书作者编写代码时采用的 IDE (integrated development environment) 是 Spyder，目的是给大家提供简洁的 Python 代码文件。

但是，建议大家采用 JupyterLab 或 Jupyter notebook 作为本系列丛书配套学习工具。

简单来说，Jupyter 集合“浏览器 + 编程 + 文档 + 绘图 + 多媒体 + 发布”众多功能与一身，非常适合探究式学习。

运行 Jupyter 无需 IDE，只需要浏览器。Jupyter 容易分块执行代码。Jupyter 支持 inline 打印结果，直接将结果图片打印在分块代码下方。Jupyter 还支持很多其他语言，比如 R 和 Julia。

使用 markdown 文档编辑功能，可以编程同时写笔记，不需要额外创建文档。Jupyter 中插入图片和视频链接都很方便。此外，还可以插入 Latex 公式。对于长文档，可以用边栏目录查找特定内容。

Jupyter 发布功能很友好，方便打印成 HTML、PDF 等格式文件。

Jupyter 也并不完美，目前尚待解决的问题有几个。Jupyter 中代码调试不方便，需要安装专门插件（比如 debugger）。Jupyter 没有 variable explorer，要么 inline 打印数据，要么将数据写到 csv 或 Excel 文件中再打开。图像结果不具有交互性，比如不能查看某个点的值，或者旋转 3D 图形，可以考虑安装 (jupyter-matplotlib)。注意，利用 Altair 或 Plotly 绘制的图像支持交互功能。对于自定义函数，目前没有快捷键直接跳转到其定义。但是，很多开发者针对这些问题都开发了插件，请大家留意。

大家可以下载安装 Anaconda，JupyterLab、Spyder、PyCharm 等常用工具都集成在 Anaconda 中。下载 Anaconda 的地址为：

◀ <https://www.anaconda.com/>

学习步骤

大家可以根据自己的偏好制定学习步骤，本书推荐如下步骤。



学完每章后，大家可以在平台上发布自己的 Jupyter 笔记，进一步听取朋友们的意见，共同进步。这样做还可以提高自己学习的动力。

意见建议

欢迎大家对本系列丛书提意见和建议，丛书专属邮箱地址为：

◀ jiang.visualize.ml@gmail.com

也欢迎大家在 B 站视频下方留言互动。

Contents

目录



0.1 本册在全套丛书的定位

“鸢尾花书”有三大板块——编程、数学、实践。数据科学、机器学习各种算法都离不开数学，因此“鸢尾花书”在数学板块着墨颇多。

本册《统计至简》是“数学三剑客”的第三本，也是最后一本。“数学”板块的第一本《数学要素》是各种数学工具的“大杂烩”，可谓数学基础。第二本《矩阵力量》专门讲解机器学习中常用的线性代数工具。本册《统计至简》则介绍机器学习和数据分析中常用的概率统计工具。

《统计至简》的核心是“多元统计”，离不开《矩阵力量》中介绍的线性代数工具。在开始本册内容学习之前，请大家务必掌握《矩阵力量》的主要内容。

在完成本册《统计至简》学习之后，我们便正式进入“实践”板块，开始《数据有道》、《机器学习》两册的探索之旅。

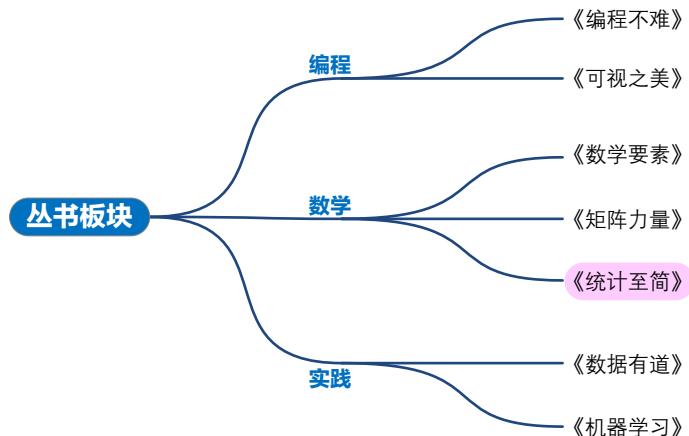


图 1. 本系列丛书板块布局

0.2 结构：7 大板块

本书可以归纳为 7 大板块——统计、概率、高斯、随机、频率派、贝叶斯派、椭圆。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>
本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

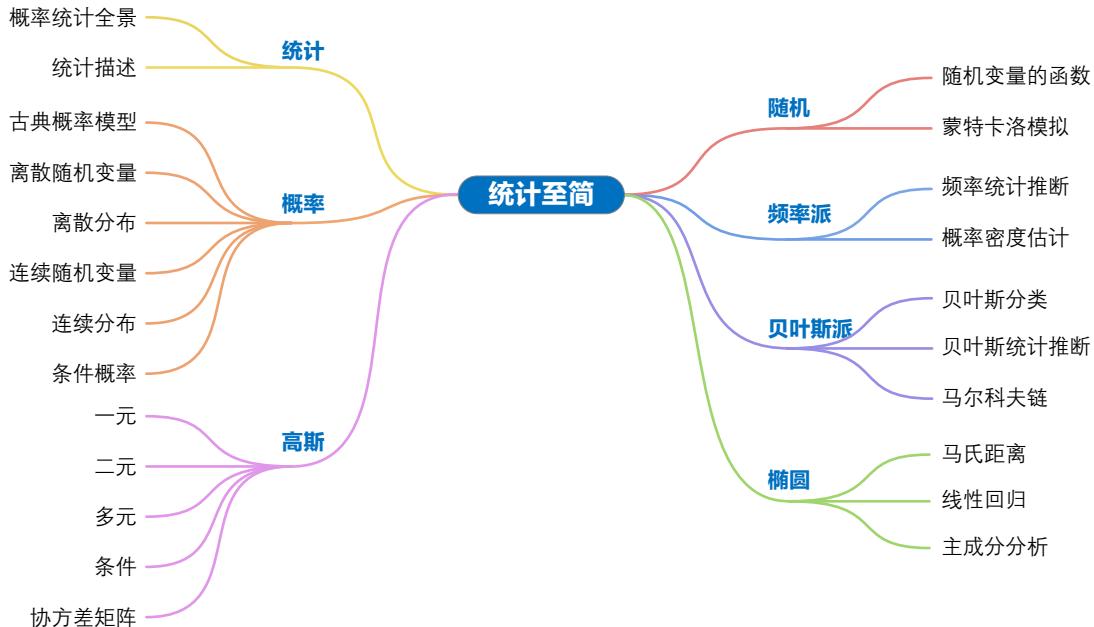


图 2. 《统计至简》板块布局

统计

本书第 1 章可能是整个“鸢尾花书”系列中“最无聊”的一章。这章首先给大家出了一个线性代数的小测验，如果顺利通过测验的话就可以开始本册内容学习。如果不顺利，建议大家回顾《矩阵力量》一册相关内容。然后，这章总结了《统计至简》重要的公式，大家可以把这些内容当成“公式手册”来看待。学习本册时或学完本册后回看参考时，大家可以试着给每个公式配图。

第 2 章介绍统计描述。这一章用图像、量化汇总等方式描述样本数据重要特征。学习这章时，建议大家回顾《矩阵力量》第 22 章。

概率

概率是统计推断的基础数学工具。“概率”这个板块将主要介绍离散、连续两大类随机变量及常见的概率分布。

第 3 章介绍古典模型，重中之重是贝叶斯定理。本书“厚”贝叶斯派，“薄”频率派，因此本书很多内容在展示贝叶斯定理的应用。希望大家从第 3 章开始就格外重视贝叶斯定理。

第 4、5 两章介绍离散随机变量、离散分布。第 6、7 两章介绍连续随机变量、连续分布。第 4、6 章特别用鸢尾花数据为例讲解随机变量，建议大家对比阅读。第 8 章特别介绍离散、连续随机变量的条件期望、条件方差。学习各种分布时，请大家格外注意它们的 PDF、CDF 形状。二项分布、多项分布、高斯分布、Dirichlet 分布这几种分布将会在本书后续发挥重要作用，希望大家留意。

学习这个板块时，请大家注意理解概率质量函数、概率密度函数无非就是对 1 (样本空间对应的概率) 的不同“切片、切块”、“切丝、切条”方式。

高斯

“高斯”是数据科学、机器学习算法中如雷贯耳的名字，大家会在回归分析、主成分分析、高斯朴素贝叶斯、高斯过程、高斯混合模型等等算法中遇到高斯分布。因此本书中高斯分布“戏份”格外吃重。

“高斯”这一板块分别介绍一元(第 9 章)、二元(第 10 章)、多元(第 11 章)、条件高斯分布(第 12 章)。几何视角是理解高斯分布的利器，大家学习这几章时，请特别注意高斯分布、椭圆、椭球之间的联系。第 13 章则介绍高斯分布中的重要成分——协方差矩阵。

这个板块，特别是在讲解多元高斯分布、协方差时，大家会看到无所不在的线性代数。

随机

第 14 章介绍随机变量的函数，请大家特别注意从几何视角理解线性变换、主成分分析。第 15 章讲解几个蒙特卡罗模拟试验，请大家掌握产生满足特定相关性的随机数的两种方法。这两种方法分别对应《矩阵力量》中介绍的 Cholesky 分解、特征值分解，建议大家在学习时回看《矩阵力量》相关内容。

频率派

本书中有关频率派的内容着墨较少，这是因为机器学习、深度学习中贝叶斯统计应用场景更为广泛。第 16 章介绍常见经典统计推断方法，请大家务必掌握最大似然估计 MLE。第 17 章讲解概率密度估计，请大家特别注意高斯核概率密度估计。

贝叶斯派

这个板块用五章介绍贝叶斯统计应用场景。

我们先从贝叶斯分类开始。第 18、19 章介绍如何利用贝叶斯定理完成鸢尾花分类，请大家掌握后验概率、证据因子、先验概率、似然概率这些概念。在贝叶斯分类算法中，优化问题可以最大化后验概率，也可以最大化联合概率，即“似然概率 \times 先验概率”。注意，《机器学习》会深入介绍“朴素贝叶斯分类”算法。

第 20、21 章讲解贝叶斯统计推断。贝叶斯统计推断把总体的模型参数看作随机变量。贝叶斯统计推断所体现出来的“学习过程”和人类认知过程极为相似，请大家注意类比。贝叶斯推断中，后验 \propto 似然 \times 先验，无疑是最重要的比例关系。此外，请大家务必掌握最大后验概率 MAP。

第 22 章简单介绍 Metropolis-Hastings 采样，并讲解如何使用 PyMC3 获得服从特定后验分布的随机数。

椭圆

本书最后一个板块可以叫“椭圆三部曲”，因为最后三章都和椭圆有关。这三章也开启了下册《数据有道》三个重要话题——数据处理、回归、降维。

第 23 章讲解马氏距离，请大家特别注意马氏距离、欧氏距离、标准化欧氏距离的区别，以及马氏距离和卡方分布的联系。

第 24 章中，我们将从最小二乘法 OLS、优化、投影、线性方程组、条件概率、最大似然估计 MLE 这几个视角讲解线性回归。这一章相当于《数学要素》第 24 章的扩展。

预告一下，《数据有道》将铺开介绍更多回归算法，比如多元回归分析、正则化、岭回归、套索回归、弹性网络回归、贝叶斯回归（最大后验估计 MAP 视角）、多项式回归、逻辑回归，以及基于主成分分析的正交回归、主元回归等算法。

第 25 章以概率统计、几何、矩阵分解、优化为视角介绍主成分分析。《数据有道》将会深入讲解主成分分析，以及典型性分析、因子分析。

0.3 特点：多元统计

《统计至简》一册最大特点就是，多元统计。

当前多数概率统计教材都侧重于“一元”，而数据科学、机器学习中处理的问题几乎都是多特征，即“多元”。从一元到多元，有一道鸿沟。能帮助我们跨越这道鸿沟的正是线性代数工具。这就是为什么一再强调大家要学好《矩阵力量》之后再开始本书学习。

概率统计是个庞杂的知识系统，本书只能选取机器学习中最常用的数学工具。“大而全”的数学公式手册范式不是本书的追求，这也就是本书书名“至简”二字的来由。本书“至简”知识体系骨架足够撑起丛书后续数据科学、机器学习内容，也方便大家进一步扩展填充。

本书“繁复”的一点是丰富的实例和可视化方案，它们可以帮助大家理解常用概率统计工具，力求让大家学透每一个公式。学习《统计至简》时，请大家注意使用几何视角，提升自己空间想象力。

阅读本册时，大家注意两个“斯”——高斯、贝叶斯。高斯分布可能是最重要的连续随机变量分布。本书把高斯分布从一元扩展到多元，关键在掌握多元高斯分布。

此外，全书每个板块几乎都有“贝叶斯定理”投下的影子。请大家务必理解条件概率、后验概率、证据因子、先验概率、似然概率在贝叶斯统计推断的应用。

“图解 + 编程 + 机器学习应用”是鸢尾花书的核心特点，本册也不例外。这套书用“编程 + 可视化”取代“习题集”。为了达到更好的学习效果，希望大家一边阅读，一边编程实践。

大多数概率统计的图书给大家的印象是公式连篇。《统计至简》为了打破这种刻板印象，尝试直接给核心公式“配图”，以强化理解。这也是本册的一个小实验，效果好的话再版的时候将推广应用到“鸢尾花书”其他分册。

此外，鸡、兔、猪这三个“小伙伴”也会来到《统计至简》客串出演，帮助大家理解复杂的概率统计概念。

“有数据的地方，必有统计！”

在《统计至简》这本书中，大家会看到微积分、线性代数、概率统计等数学工具“济济一堂”，但是没有丝毫的违和感！

下面，我们就开始“数学三剑客”的收官之旅！

1

Landscape of Statistics and Probability

本书概率统计全景

公式连篇，可能是“鸢尾花书”最枯燥无味的一章



概率论作为数学学科，可以且应该从公理开始建设，和几何、代数的思路一样。

The theory of probability as mathematical discipline can and should be developed from axioms in exactly the same way as Geometry and Algebra.

—— 安德雷·柯尔莫哥洛夫 (Andrey Kolmogorov) | 概率论公理化之父 | 1903 ~ 1987

1.1

必备数学工具：一个线性代数小测验

本书前文提到，《统计至简》一册的核心特点是——多元。《矩阵力量》中介绍的线性代数工具是本书核心数学工具。因此，在开始本书阅读之前，请大家完成本节这个小测验。

如果大家能够轻松完成这个测验，欢迎大家开始本书后续内容学习；否则，建议大家重温《矩阵力量》中相关数学工具。

数据矩阵

给定数据矩阵 X ，如何求其质心、中心化数据、标准化数据、格拉姆矩阵、协方差矩阵、相关系数矩阵？

协方差矩阵

给定 2×2 协方差矩阵 Σ ：

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (\text{test.1})$$

什么条件下 Σ 是正定矩阵？

定义如下二元函数：

$$f(x_1, x_2) = \mathbf{x}^\top \Sigma \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^\top \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad (\text{test.2})$$

相关性系数 $\rho_{1,2}$ 的取值范围是什么？上述二元函数的图像是什么？

当 σ_1 和 σ_2 均为 1 时，这个二元函数等高线形状随 $\rho_{1,2}$ 如何变化？

Cholesky 分解

对协方差矩阵 Σ 进行 Cholesky 分解：

$$\Sigma = \mathbf{R}^\top \mathbf{R} \quad (\text{test.3})$$

矩阵 Σ 能进行 Cholesky 分解的前提是什么？

上三角矩阵 \mathbf{R} 的特点是什么？如何从几何角度理解 \mathbf{R} ？

特征值分解

对 Σ 特征值分解：

$$\Sigma = V \Lambda V^T \quad (\text{test.4})$$

等式右侧第二个矩阵 V 对应转置运算，为什么？

矩阵 V 有什么特殊性质？如何从向量空间角度理解 V ？

矩阵 Λ 有什么特殊性质？什么条件下， Σ 特征值中有 0？

如果把 V 写成 $[v_1, v_2]$ ，上式可以如何展开？

将 (test.4) 写成：

$$V^T \Sigma V = \Lambda \quad (\text{test.5})$$

把 V 写成 $[v_1, v_2]$ ，上式如何展开？

几何角度来看，上式代表什么？

奇异值分解

奇异值分解有哪四种类型？每种类型之间存在怎样的关系？

数据矩阵 X 奇异值分解可以获得其奇异值 s_j ，对 X 的格拉姆矩阵 G 特征值分解可以得到特征值 $\lambda_{G,j}$ 。奇异值 s_j 和特征值 $\lambda_{G,j}$ 存在怎样的量化关系？

对 X 的协方差矩阵 Σ 特征值分解可以得到特征值 λ_j 。奇异值 s_j 和特征值 λ_j 又存在怎样的量化关系？

奇异值分解和向量四个空间有怎样联系？

多元高斯分布

多元正态分布的概率密度函数 PDF 为：

$$f_x(x) = \frac{\exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \quad (\text{test.6})$$

$(x - \mu)^T \Sigma^{-1} (x - \mu)$ 的含义是什么？

$(2\pi)^{\frac{D}{2}}$ 的作用是什么？ $|\Sigma|^{\frac{1}{2}}$ 的含义是什么？

什么情况下，上式不成立？

马氏距离的定义是什么？马氏距离和欧氏距离差别是什么？

测验题目到此结束。



本书不就上述题目给出具体答案，所有答案都在《矩阵力量》一册，请大家自行查阅。

本章下面先用数学手册、备忘录这种范式罗列本书中 100 个核心公式，每一节对应本书一个板块。而本章之后，我们就用丰富的图形给这些公式以色彩和温度。

1.2 统计描述

给定随机变量 X 的 n 个样本 $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ ， X 的样本均值为：

$$\mu_X = \frac{1}{n} \left(\sum_{i=1}^n x^{(i)} \right) = \frac{x^{(1)} + x^{(2)} + x^{(3)} + \dots + x^{(n)}}{n} \quad (1)$$

X 的样本方差为：

$$\text{var}(X) = \sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)^2 \quad (2)$$

X 的样本标准差为：

$$\sigma_X = \text{std}(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)^2} \quad (3)$$

对于样本数据，随机变量 X 和 Y 的协方差为：

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)(y^{(i)} - \mu_Y) \quad (4)$$

对于样本数据，随机变量 X 和 Y 的相关性系数为：

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (5)$$

⚠ 注意，除非特殊说明，本书一般不从符号上区分总体、样本的均值、方差、标准差等。

1.3 概率

古典概率模型

设样本空间 Ω 由 n 个等可能事件构成，事件 A 的概率为：

$$\Pr(A) = \frac{n_A}{n} \quad (6)$$

其中， n_A 为含于事件 A 的试验结果数量。

A 和 B 为样本空间 Ω 中的两个事件，其中 $\Pr(B) > 0$ 。那么，事件 B 发生的条件下事件 A 发生的条件概率为：

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)} \quad (7)$$

其中， $\Pr(A, B)$ 为 A 和 B 事件的联合概率， $\Pr(B)$ 也叫 B 事件边缘概率。

类似地，如果 $\Pr(A) > 0$ ，事件 A 发生的条件下事件 B 发生的条件概率为：

$$\Pr(B|A) = \frac{\Pr(A, B)}{\Pr(A)} \quad (8)$$

贝叶斯定理为：

$$\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A) = \Pr(A, B) \quad (9)$$

假设 A_1, A_2, \dots, A_n 互不相容，形成对样本空间 Ω 的分割。 $\Pr(A_i) > 0$ ，对于空间 Ω 中任意事件 B ，全概率定理为：

$$\Pr(B) = \sum_{i=1}^n \Pr(A_i, B) \quad (10)$$

如果事件 A 和事件 B 独立，则：

$$\begin{aligned} \Pr(A|B) &= \Pr(A) \\ \Pr(B|A) &= \Pr(B) \\ \Pr(A, B) &= \Pr(A)\Pr(B) \end{aligned} \quad (11)$$

如果事件 A 和事件 B 在 C 发生条件下条件独立，则：

$$\Pr(A, B|C) = \Pr(A|C) \cdot \Pr(B|C) \quad (12)$$

离散随机变量

离散随机变量 X 的概率质量函数满足：

$$\sum_x p_X(x) = 1, \quad 0 \leq p_X(x) \leq 1 \quad (13)$$

离散随机变量 X 的期望值为：

$$E(X) = \sum_x x \cdot p_X(x) \quad (14)$$

离散随机变量 X 的方差为：

$$\text{var}(X) = \sum_x (x - \text{E}(X))^2 \cdot p_X(x) \quad (15)$$

二元离散随机变量 (X, Y) 的概率质量函数满足：

$$\sum_x \sum_y p_{X,Y}(x,y) = 1, \quad 0 \leq p_{X,Y}(x,y) \leq 1 \quad (16)$$

(X, Y) 的协方差定义为：

$$\begin{aligned} \text{cov}(X, Y) &= \text{E}((X - \text{E}(X))(Y - \text{E}(Y))) \\ &= \sum_x \sum_y p_{X,Y}(x,y)(x - \text{E}(X))(y - \text{E}(Y)) \end{aligned} \quad (17)$$

边缘概率 $p_X(x)$ 为：

$$p_X(x) = \sum_y p_{X,Y}(x,y) \quad (18)$$

边缘概率 $p_Y(y)$ 为：

$$p_Y(y) = \sum_x p_{X,Y}(x,y) \quad (19)$$

在给定事件 $\{Y=y\}$ 条件下， $p_Y(y) > 0$ ，事件 $\{X=x\}$ 发生的条件概率质量函数 $p_{X|Y}(x|y)$ 为：

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \quad (20)$$

$p_{X|Y}(x|y)$ 对 x 求和等于 1：

$$\sum_x p_{X|Y}(x|y) = 1 \quad (21)$$

在给定事件 $\{X=x\}$ 条件下， $p_X(x) > 0$ ，事件 $\{Y=y\}$ 发生的条件概率质量函数 $p_{Y|X}(y|x)$ 为：

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \quad (22)$$

$p_{Y|X}(y|x)$ 对 y 求和等于 1：

$$\sum_y p_{Y|X}(y|x) = 1 \quad (23)$$

如果离散随机变量 X 和 Y 独立，则：

$$\begin{aligned} p_{X|Y}(x|y) &= p_X(x) \\ p_{Y|X}(y|x) &= p_Y(y) \\ p_{X,Y}(x,y) &= p_Y(y) \cdot p_X(x) \end{aligned} \quad (24)$$

离散分布

$[a, b]$ 上离散均匀分布的概率质量函数为：

$$p_x(x) = \frac{1}{b-a+1}, \quad x = a, a+1, \dots, b-1, b \quad (25)$$

伯努利分布的概率质量函数为：

$$p_x(x) = p^x (1-p)^{1-x} \quad x \in \{0, 1\} \quad (26)$$

其中， p 的取值范围为 $[0, 1]$ 。

二项分布的概率质量函数为：

$$p_x(x) = C_n^x p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n \quad (27)$$

多项分布的概率质量函数为：

$$p_{x_1, \dots, x_K}(x_1, \dots, x_K; n, p_1, \dots, p_K) = \begin{cases} \frac{n!}{(x_1!) \times (x_2!) \cdots \times (x_K!)} \times p_1^{x_1} \times \cdots \times p_K^{x_K} & \text{when } \sum_{i=1}^K x_i = n \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

其中 $x_i (i = 1, 2, \dots, K)$ 为非负整数； p_i 取值范围为 $(0, 1)$ ，且 $\sum_{i=1}^K p_i = 1$ 。

泊松分布的概率质量函数为：

$$p_x(x) = \frac{\exp(-\lambda) \lambda^x}{x!}, \quad x = 0, 1, 2, \dots \quad (29)$$

其中， λ 大于 0。 λ 既是期望值，也是方差。

连续随机变量

连续随机变量 X 的概率密度函数满足：

$$\int_{-\infty}^{+\infty} f_X(x) dx = 1, \quad f_X(x) \geq 0 \quad (30)$$

连续随机变量 X 期望为：

$$E(X) = \int_x x \cdot f_X(x) dx \quad (31)$$

连续随机变量 X 方差为：

$$\text{var}(X) = E[(X - E(X))^2] = \int_x (x - E(X))^2 \cdot f_X(x) dx \quad (32)$$

给定 (X, Y) 的联合概率分布 $f_{X,Y}(x,y)$ ， X 的边缘概率密度函数 $f_X(x)$ 为：

$$f_X(x) = \int_y f_{X,Y}(x,y) dy \quad (33)$$

连续随机变量 Y 的边缘概率密度函数 $f_Y(y)$ 为：

$$f_Y(y) = \int_x f_{X,Y}(x,y) dx \quad (34)$$

在给定 $Y=y$ 条件下，且 $f_Y(y) > 0$ ，条件概率密度函数 $f_{X|Y}(x|y)$ 为：

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad (35)$$

给定 $X=x$ 条件下，且 $f_X(x) > 0$ ，条件概率密度函数 $f_{Y|X}(y|x)$ 为：

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \quad (36)$$

利用贝叶斯定理，联合概率 $f_{X,Y}(x,y)$ 为：

$$f_{X,Y}(x,y) = f_{X|Y}(x|y) f_Y(y) = f_{Y|X}(y|x) f_X(x) \quad (37)$$

如果连续随机变量 X 和 Y 独立，则：

$$\begin{aligned} f_{X|Y}(x|y) &= f_X(x) \\ f_{Y|X}(y|x) &= f_Y(y) \\ f_{X,Y}(x,y) &= f_X(x) f_Y(y) \end{aligned} \quad (38)$$

连续分布

区间 $[a, b]$ 的连续均匀分布概率密度函数为：

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \quad (39)$$

一元学生 t -分布的概率密度函数为：

$$f_X(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad (40)$$

其中， ν 大于 0。

指数分布的概率密度函数为：

$$f_x(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (41)$$

其中， λ 大于 0。

Beta(α, β) 分布的概率密度函数为：

$$f_x(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (42)$$

其中， α 和 β 均大于 0。这个 PDF 也可以写成：

$$f_x(x; \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (43)$$

其中，Beta 函数 $B(\alpha, \beta)$ 为：

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (44)$$

Dirichlet 分布概率密度函数为：

$$f_{x_1, \dots, x_K}(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} \prod_{i=1}^K x_i^{\alpha_i-1}, \quad \sum_{i=1}^K x_i = 1 \quad (45)$$

其中， α_i 大于 0。

⚠ 注意，对于 Dirichlet 分布，本书后续常用变量 θ 代替 x 。

Beta 函数 $B(\alpha_1, \dots, \alpha_K)$ 为：

$$B(\alpha_1, \dots, \alpha_K) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i\right)} \quad (46)$$

条件概率

如果 X 和 Y 均为离散随机变量，给定 $X=x$ 条件下， Y 的条件期望 $E(Y|X=x)$ 为：

$$E(Y|X=x) = \sum_y y \cdot p_{Y|X}(y|x) \quad (47)$$

$E(Y)$ 的全期望定理为：

$$E(Y) = E(E(Y|X)) = \sum_x E(Y|X=x) \cdot p_X(x) \quad (48)$$

给定 $Y=y$ 条件下， X 的条件期望 $E(X|Y=y)$ 定义为：

$$\mathbb{E}(X|Y=y) = \sum_x x \cdot p_{x|y}(x|y) \quad (49)$$

$\mathbb{E}(X)$ 的全期望定理为：

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) = \sum_y \mathbb{E}(X|Y=y) \cdot p_Y(y) \quad (50)$$

给定 $X=x$ 条件下， Y 的条件方差 $\text{var}(Y|X=x)$ 为：

$$\text{var}(Y|X=x) = \sum_y (y - \mathbb{E}(Y|X=x))^2 \cdot p_{y|x}(y|x) \quad (51)$$

给定 $Y=y$ 条件下， X 的条件方差 $\text{var}(X|Y=y)$ 为：

$$\text{var}(X|Y=y) = \sum_x (x - \mathbb{E}(X|Y=y))^2 \cdot p_{x|y}(x|y) \quad (52)$$

对于 $\text{var}(Y)$ ，全方差定理为：

$$\text{var}(Y) = \mathbb{E}(\text{var}(Y|X)) + \text{var}(\mathbb{E}(Y|X)) \quad (53)$$

对于 $\text{var}(X)$ ，全方差定理为：

$$\text{var}(X) = \mathbb{E}(\text{var}(X|Y)) + \text{var}(\mathbb{E}(X|Y)) \quad (54)$$

如果 X 和 Y 均为连续随机变量，在给定 $X=x$ 条件下，条件期望 $\mathbb{E}(Y|X=x)$ 为：

$$\mathbb{E}(Y|X=x) = \int_y y \cdot f_{y|x}(y|x) dy \quad (55)$$

条件方差 $\text{var}(Y|X=x)$ 为：

$$\text{var}(Y|X=x) = \int_y (y - \mathbb{E}(Y|X=x))^2 \cdot f_{y|x}(y|x) dy \quad (56)$$

在给定 $Y=y$ 条件下，条件期望 $\mathbb{E}(X|Y=y)$ 为：

$$\mathbb{E}(X|Y=y) = \int_x x \cdot f_{x|y}(x|y) dx \quad (57)$$

条件方差 $\text{var}(X|Y=y)$ 定义为：

$$\text{var}(X|Y=y) = \int_x (x - \mathbb{E}(X|Y=y))^2 \cdot f_{x|y}(x|y) dx \quad (58)$$

1.4 高斯

一元高斯分布

一元高斯分布的概率密度函数为：

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (59)$$

标准正态分布的概率密度函数为：

$$f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (60)$$

二元高斯分布

如果 (X, Y) 服从二元高斯分布，且相关性系数不为 ± 1 ， (X, Y) 的概率密度函数为：

$$f_{x,y}(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \times \exp\left(-\frac{1}{2}\left(\frac{1}{1-\rho_{x,y}^2}\left(\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho_{x,y}\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right)\right)\right) \quad (61)$$

X 的边缘概率密度函数为：

$$f_x(x) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_x}{\sigma_x}\right)^2\right) \quad (62)$$

Y 的边缘概率密度函数为：

$$f_y(y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right) \quad (63)$$

多元高斯分布

多元高斯分布的概率密度函数为：

$$f_x(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \quad (64)$$

其中，协方差矩阵 Σ 为正定矩阵。

条件高斯分布

如果 (X, Y) 服从二元高斯分布，且相关性系数不为 ± 1 ， $f_{Y|X}(y|x)$ 为：

$$f_{Y|X}(y|x) = \frac{1}{\sigma_Y \sqrt{1-\rho_{X,Y}^2} \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{y - \left(\mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_X)\right)}{\sigma_Y \sqrt{1-\rho_{X,Y}^2}} \right)^2\right) \quad (65)$$

条件期望 $E(Y|X=x)$ 为：

$$E(Y|X=x) = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_X) \quad (66)$$

条件方差 $\text{var}(Y|X=x)$ 为：

$$\text{var}(Y|X=x) = (1 - \rho_{X,Y}^2) \sigma_Y^2 \quad (67)$$

如果随机变量向量 χ 和 γ 服从多元高斯分布：

$$\begin{bmatrix} \chi \\ \gamma \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_\chi \\ \mu_\gamma \end{bmatrix}, \begin{bmatrix} \Sigma_{\chi\chi} & \Sigma_{\chi\gamma} \\ \Sigma_{\gamma\chi} & \Sigma_{\gamma\gamma} \end{bmatrix}\right) \quad (68)$$

其中，

$$\chi = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}, \quad \gamma = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{bmatrix} \quad (69)$$

给定 $\chi = x$ 的条件下， γ 服从如下多元高斯分布：

$$\{\gamma|\chi=x\} \sim N\left(\underbrace{\Sigma_{\gamma\chi} \Sigma_{\chi\chi}^{-1} (x - \mu_\chi) + \mu_\gamma}_{\text{Expectation}}, \underbrace{\Sigma_{\gamma\gamma} - \Sigma_{\gamma\chi} \Sigma_{\chi\chi}^{-1} \Sigma_{\chi\gamma}}_{\text{Variance}}\right) \quad (70)$$

给定 $\chi = x$ 的条件下 γ 的条件期望为：

$$E(\gamma|\chi=x) = \mu_{\gamma|\chi=x} = \Sigma_{\gamma\chi} \Sigma_{\chi\chi}^{-1} (x - \mu_\chi) + \mu_\gamma \quad (71)$$

协方差矩阵

随机变量向量 χ 的协方差矩阵为：

$$\begin{aligned} \text{var}(\chi) &= \text{cov}(\chi, \chi) = E[(\chi - E(\chi))(\chi - E(\chi))^T] \\ &= E(\chi \chi^T) - E(\chi) E(\chi)^T \end{aligned} \quad (72)$$

样本数据矩阵 X 的协方差矩阵 Σ 为：

$$\Sigma = \frac{(\mathbf{X} - \mathbb{E}(\mathbf{X}))^\top (\mathbf{X} - \mathbb{E}(\mathbf{X}))}{n-1} \quad (73)$$

合并协方差矩阵为：

$$\Sigma_{\text{pooled}} = \frac{1}{\sum_{k=1}^K (n_k - 1)} \sum_{k=1}^K (n_k - 1) \Sigma_k = \frac{1}{n - K} \sum_{k=1}^K (n_k - 1) \Sigma_k \quad (74)$$

其中， $\sum_{k=1}^K n_k = n$ 。

1.5 随机

随机变量的函数

如果 Y 和二元随机变量 (X_1, X_2) 存在如下关系：

$$Y = aX_1 + bX_2 = [a \ b] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (75)$$

Y 的期望、方差为：

$$\mathbb{E}(Y) = [a \ b] \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{bmatrix}, \quad \text{var}(Y) = [a \ b] \underbrace{\begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \text{var}(X_2) \end{bmatrix}}_{\Sigma} \begin{bmatrix} a \\ b \end{bmatrix} \quad (76)$$

如果 $\chi = [X_1, X_2, \dots, X_D]^\top$ 服从 $N(\mu_\chi, \Sigma_\chi)$ ， χ 在单位向量 v 方向上投影得到 Y ：

$$Y = v^\top \chi \quad (77)$$

Y 的期望、方差为：

$$\begin{aligned} \mathbb{E}(Y) &= v^\top \mu_\chi \\ \text{var}(Y) &= v^\top \Sigma_\chi v \end{aligned} \quad (78)$$

χ 在规范正交系 V 投影得到 γ ：

$$\gamma = V^\top \chi \quad (79)$$

γ 的期望、协方差矩阵为：

$$\begin{aligned} \mathbb{E}(\gamma) &= V^\top \mu_\chi \\ \text{var}(\gamma) &= V^\top \Sigma_\chi V \end{aligned} \quad (80)$$

1.6 频率派

频率派统计推断

随机变量 X_1, X_2, \dots, X_n 独立同分布。 $X_k (k = 1, 2, \dots, n)$ 的期望和方差为：

$$\mathbb{E}(X_k) = \mu, \quad \text{var}(X_k) = \sigma^2 \quad (81)$$

这 n 个随机变量的平均值 \bar{X} 近似服从如下正态分布：

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (82)$$

最大似然估计的优化问题为：

$$\hat{\theta}_{MLE} = \arg \max_{\theta} \prod_{i=1}^n f_{X_i}(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \ln f_{X_i}(x_i; \theta) \quad (83)$$

概率密度估计

概率密度估计函数为：

$$\hat{f}_X(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x^{(i)}) = \frac{1}{n} \frac{1}{h} \sum_{i=1}^n K\left(\frac{x - x^{(i)}}{h}\right), \quad -\infty < x < +\infty \quad (84)$$

核函数 $K(x)$ 满足两个重要条件：(1) 对称性；(2) 面积为 1：

$$\begin{aligned} K(x) &= K(-x) \\ \int_{-\infty}^{+\infty} K(x) dx &= \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x}{h}\right) dx = 1 \end{aligned} \quad (85)$$

1.7 贝叶斯派

贝叶斯分类

利用贝叶斯定理分类：

$$f_{Y|X}(C_k | x) = \frac{f_{X|Y}(x | C_k) p_Y(C_k)}{f_X(x)} \quad (86)$$

$f_{Y|X}(C_k | x)$ 叫后验概率，又叫成员值。

$f_X(x)$ 为证据因子，也叫证据。

$p_Y(C_k)$ 为先验概率，表达样本集合中 C_k 类样本占比。

$f_{X|Y}(x|C_k)$ 为似然概率。

贝叶斯分类优化问题：

$$\hat{y} = \arg \max_{C_k} f_{Y|X}(C_k|x) = \arg \max_{C_k} f_{X|Y}(x|C_k) p_Y(C_k) \quad (87)$$

其中， $k = 1, 2, \dots, K$ 。

贝叶斯统计推断

模型参数的后验分布为：

$$f_{\Theta|X}(\theta|x) = \frac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int_{\theta} f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)d\theta} \quad (88)$$

后验 \propto 似然 \times 先验，最大化后验估计的优化问题等价于：

$$\hat{\theta}_{MAP} = \arg \max_{\theta} f_{\Theta|X}(\theta|x) = \arg \max_{\theta} f_{X|\Theta}(x|\theta)f_{\Theta}(\theta) \quad (89)$$

1.8 椭圆三部曲

马氏距离

马氏距离的定义为：

$$d = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} \quad (90)$$

D 维马氏距离的平方则服从自由度为 D 的卡方分布：

$$d^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \sim \chi^2_{(df=D)} \quad (91)$$

线性回归

多元线性回归可以写成超定方程组：

$$y = Xb \quad (92)$$

如果 $X^T X$ 可逆，则 b 为：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>
本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (93)$$

主成分分析

⚠ 注意，这部分公式实际上来自《矩阵力量》；此外，我们将会在《数据有道》也会用到这些公式。

对原始矩阵 \mathbf{X} 进行经济型 SVD 分解：

$$\mathbf{X} = \mathbf{U}_x \mathbf{S}_x \mathbf{V}_x^T \quad (94)$$

其中， \mathbf{S}_x 为对角方阵。

利用 (94)， \mathbf{X} 的格拉姆矩阵可以展开为：

$$\mathbf{G} = \mathbf{V}_x \mathbf{S}_x^2 \mathbf{V}_x^T \quad (95)$$

上式便是格拉姆 \mathbf{G} 的特征值分解。

对中心化数据矩阵 \mathbf{X}_c 经济型 SVD 分解：

$$\mathbf{X}_c = \mathbf{U}_c \mathbf{S}_c \mathbf{V}_c^T \quad (96)$$

而协方差矩阵 Σ 则可以写成：

$$\Sigma = \mathbf{V}_c \frac{\mathbf{S}_c^2}{n-1} \mathbf{V}_c^T \quad (97)$$

相信大家在上式中能够看到协方差矩阵 Σ 的特征值分解。请大家注意 (96) 中奇异值和 (97) 中特征值关系：

$$\lambda_{c-j} = \frac{s_{c-j}^2}{n-1} \quad (98)$$

同样，对标准化数据矩阵 \mathbf{Z}_x 进行经济型 SVD 分解：

$$\mathbf{Z}_x = \mathbf{U}_z \mathbf{S}_z \mathbf{V}_z^T \quad (99)$$

相关性系数矩阵 \mathbf{P} 则可以写成：

$$\mathbf{P} = \mathbf{V}_z \frac{\mathbf{S}_z^2}{n-1} \mathbf{V}_z^T \quad (100)$$

上式相当于对 \mathbf{P} 特征值分解。



学完本册《统计至简》后，再回过头来看本章罗列的这些公式时，希望大家看到的不再是冷冰冰的符号，而是一幅幅彩色的图像。



Descriptive Statistics

2 统计描述

用图形和汇总统计量描述样本数据



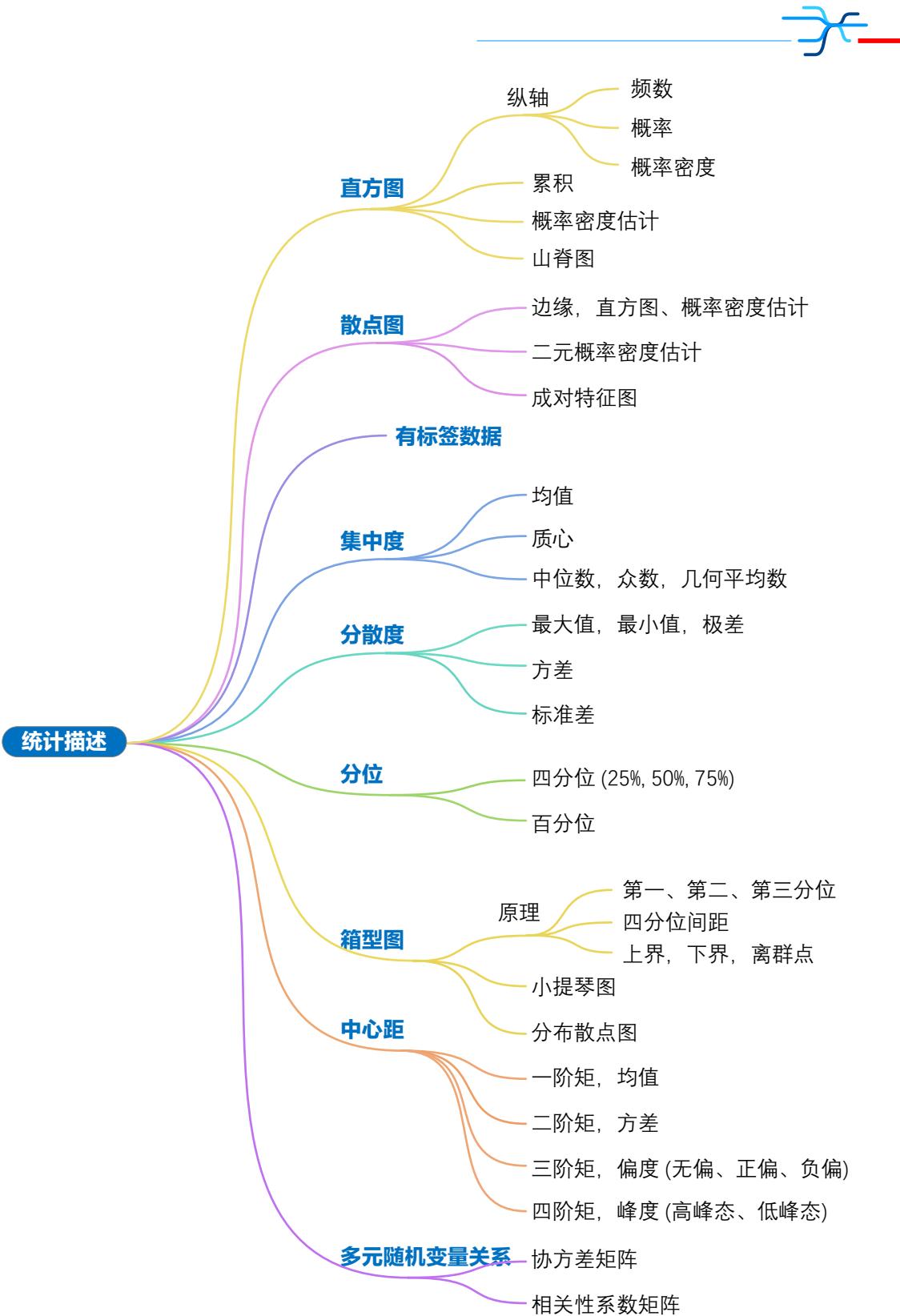
统计学是科学的语法。

Statistics is the grammar of science.

—— 卡尔·皮尔逊 (Karl Pearson) | 英国数学家 | 1857 ~ 1936



- ▶ `joyplot()` 绘制山脊图
- ▶ `numpy.percentile()` 计算百分位
- ▶ `pandas.plotting.parallel_coordinates()` 绘制平行坐标图
- ▶ `seaborn.boxplot()` 绘制箱型图
- ▶ `seaborn.heatmap()` 绘制热图
- ▶ `seaborn.histplot()` 绘制频数/概率/概率密度直方图
- ▶ `seaborn.jointplot()` 绘制联合分布和边缘分布
- ▶ `seaborn.kdeplot()` 绘制 KDE 核概率密度估计曲线
- ▶ `seaborn.lineplot()` 绘制线图
- ▶ `seaborn.lmplot()` 绘制线性回归图像
- ▶ `seaborn.pairplot()` 绘制成对分析图
- ▶ `seaborn.swarmplot()` 绘制蜂群图
- ▶ `seaborn.violinplot()` 绘制小提琴图



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

2.1 统计两大工具：描述、推断

如图 1 所示，本书中统计版图可以分为两大板块——描述、推断。

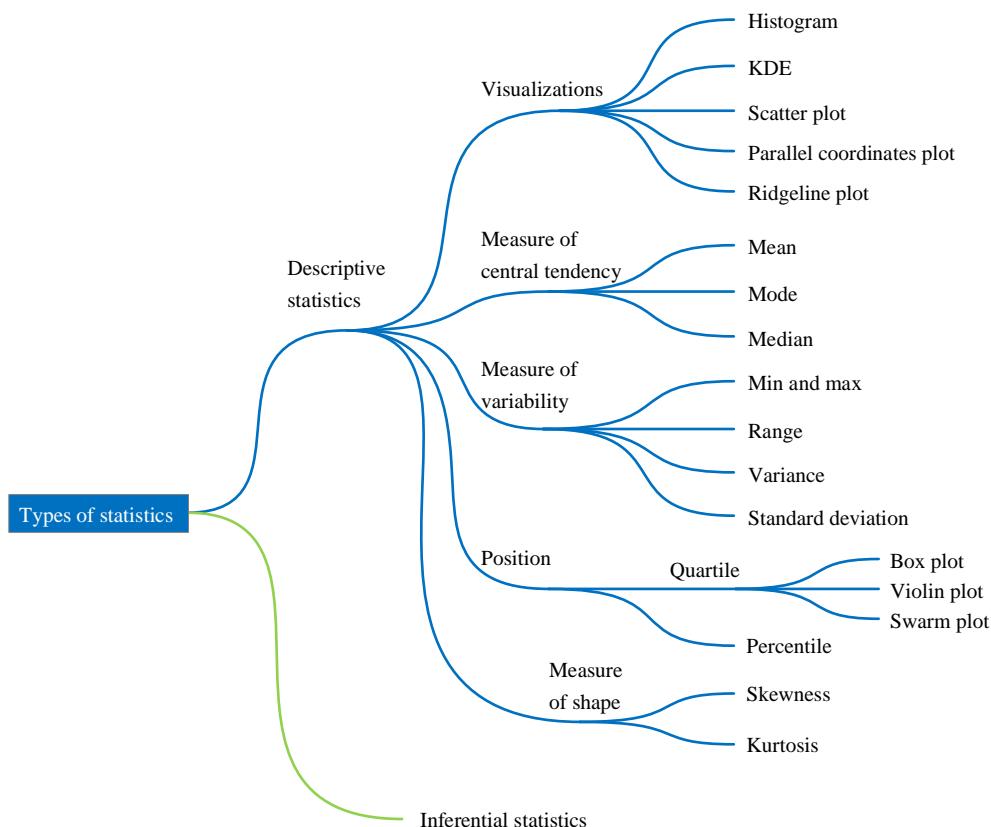


图 1. 两大类统计工具的分类

统计描述 (descriptive statistics) 是指对数据进行整体性的描述和概括，以了解数据的特征和结构。统计描述旨在通过一些表格、图像、量化汇总来呈现数据的基本特征，比如中心趋势、离散程度、分布形态等。统计描述通常是数据分析的第一步，可以帮助我们了解数据的基本情况，判断数据的可靠性、准确性和有效性。

统计推断 (statistical inference) 根据样本数据推断总体特征。统计推断是在对样本数据统计描述的基础上，对总体未知量化特征做出概率形式的推断。显然，统计推断的数学基础工具就是概率论。本书后续概率、高斯、随机这三个板块介绍概率论这个工具箱中常用工具。之后，我们将用频率派、贝叶斯派两个板块介绍统计推断。

请大家学习这一章时，重温《矩阵力量》第 22 章，回顾如何从线性代数视角看各种统计量。

本章主要介绍统计描述。常见的统计描述方法包括：

- ▶ 统计图表：可视化数据分布情况和异常值，比如直方图、箱线图、散点图等。
- ▶ 中心趋势：比如均值、中位数和众数，量化数据的集中程度。
- ▶ 离散程度：比如极差、方差、标准差、四分位数，描述数据的分散程度。
- ▶ 分布形态：比如偏度、峰度，分析数据的分布形态。
- ▶ 协同关系：包括协方差矩阵、相关性系数矩阵，量化多元随机变量之间的关系。

下面，我们开始本章学习。

2.2 直方图：单特征数据分布

鸢尾花花萼长度的数据看上去杂乱无章，我们可以利用一些统计工具来分析这组数据，比如直方图。直方图 (histogram) 由一系列矩形组成，它的横轴为组距，纵轴可以为频数 (frequency, count)、概率 (probability)、概率密度 (probability density 或 density)。

直方图可视化样本分布情况，同时展示均值、众数、中位数的大致位置以及标准差宽度等。直方图也可以用来判断数据是否存在离群值 (outlier)。



《数据有道》一册将专门讲解判断离群值的常用算法。

图 2 所示为鸢尾花花萼长度数据直方图。直方图通常将样本数据分成若干个连续的区间，也称为“箱子”或“组”。直方图中矩形的纵轴高度可以对应频数、概率或概率密度。

⚠ 再次强调，一般情况，直方图的纵轴有三个选择——频数、概率和概率密度。

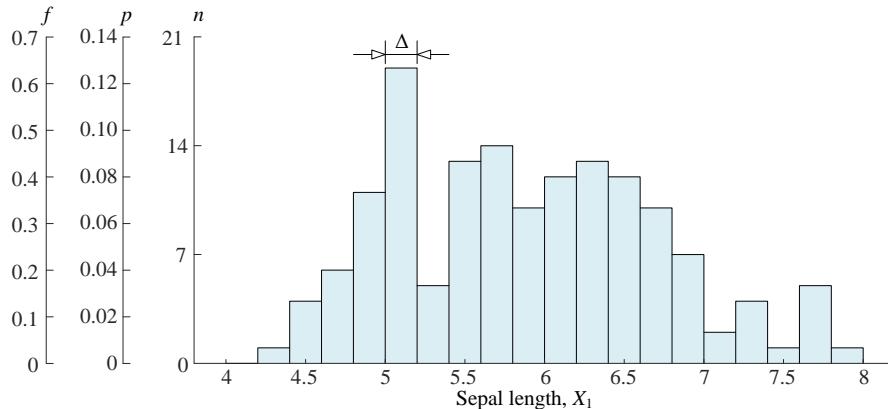


图 2. 鸢尾花花萼长度，频数、概率和概率密度的关系

下面聊聊频数、概率和概率密度分别是什么。

区间

花萼长度的最小值和最大值落在 [4, 8] 这个区间。如图 3 所示，将这个区间等分为 20 个区间。区间个数称为组数，记做 M 。每个区间对应的宽度叫做组距，记做 Δ 。本例中组数 $M = 20$ ，组距 $\Delta = 0.2 \text{ cm} = 4 \text{ cm}/20$ 。

图 3 第一列给出的是每个组距所在的区间。



鸢尾花书《数学要素》第 6 章介绍过各种不同区间类型，建议大家回顾。

区间	频数 n	累积频数 $\text{cumsum}(n)$	概率 p	累积概率 $\text{cumsum}(p)$	概率密度 f
[4.2, 4.4)	1	1	0.007	0.007	0.033
[4.4, 4.6)	4	5	0.027	0.033	0.133
[4.6, 4.8)	6	11	0.040	0.073	0.200
[4.8, 5.0)	11	22	0.073	0.147	0.367
[5.0, 5.2)	19	41	0.127	0.273	0.633
[5.2, 5.4)	5	46	0.033	0.307	0.167
[5.4, 5.6)	13	59	0.087	0.393	0.433
[5.6, 5.8)	14	73	0.093	0.487	0.467
[5.8, 6.0)	10	83	0.067	0.553	0.333
[6.0, 6.2)	12	95	0.080	0.633	0.400
[6.2, 6.4)	13	108	0.087	0.720	0.433
[6.4, 6.6)	12	120	0.080	0.800	0.400
[6.6, 6.8)	10	130	0.067	0.867	0.333
[6.8, 7.0)	7	137	0.047	0.913	0.233
[7.0, 7.2)	2	139	0.013	0.927	0.067
[7.2, 7.4)	4	143	0.027	0.953	0.133
[7.4, 7.6)	1	144	0.007	0.960	0.033
[7.6, 7.8)	5	149	0.033	0.993	0.167
[7.8, 8.0]	1	150	0.007	1.000	0.033

图 3. 鸢尾花花萼长度直方图数据

⚠ 注意，一般情况，除了最后一个区间之外，其他区间包含左侧端点，不含右侧端点，即左闭右开区间。最后一个区间为闭区间。大家已经看到图 3 最后一个区间 [7.8, 8.0] 为闭区间，其他区间均为左闭右开。

频数

频数，也叫次数，是指在一定范围内样本数据的数量。显然，频数为非负整数。如图 3 所示，落在 4.2 ~ 4.4 这个区间的样本只有 1 个。而落在 5 ~ 5.2 这个区间的样本多达 19 个。

数出落在第 i 个区间内的样本数量，定义为频数 n_i 。图 3 第二列给出的就是频数。

显然，所有频数 n_i 之和为样本总数 n ：

$$\sum_{i=1}^M n_i = n \quad (1)$$

概率

频数 n_i 除以样本总数 n 的结果做概率 p_i :

$$p_i = \frac{n_i}{n} \quad (2)$$

图 3 第四列对应概率。容易知道概率值 p_i 的取值范围 $[0, 1]$ 。概率值代表“可能性”。

直方图的纵轴为概率时，直方图也叫归一化直方图。这是因为所有区间概率 p_i 之和为 1:

$$\sum_{i=1}^M p_i = \sum_{i=1}^M \frac{n_i}{n} = \frac{n_1 + n_2 + \dots + n_M}{n} = 1 \quad (3)$$

概率密度

概率 p_i 除以组距 Δ 得到的是**概率密度** (probability density) f_i :

$$f_i = \frac{p_i}{\Delta} = \frac{n_i}{n\Delta} \quad (4)$$

纵轴为概率密度的直方图，所有矩形面积之和为 1:

$$\sum_{i=1}^M f_i \Delta = \sum_{i=1}^M \frac{p_i}{\Delta} \Delta = \sum_{i=1}^M \frac{n_i}{n} = 1 \quad (5)$$

观察图 3，我们可以发现频数、概率、概率密度这三个值成正比关系。不同的是，看频数、概率时，我们在乎的是直方图矩形高度；而看概率密度时，我们关注的是矩形面积。

⚠ 注意，概率密度不是概率；但是，概率密度本身也反映数据分布的疏密情况。

累积

图 3 中第三和第五列分别为**累积频数** (cumulative frequency) 和**累积概率** (cumulative probability)。累积频数就是将从小到大各区间的频数逐个累加起来，累积频数的最后一个值是样本总数。

类似地，我们可以得到累积概率，累积概率的最后一个值为 1。

绘制直方图

图 4 所示为利用 seaborn.histplot() 绘制的鸢尾花四个量化特征数据直方图，纵轴为频数。直方图的形状可以反映数据的分布情况，比如对称分布、左偏分布、右偏分布等。直方图可以通过调整箱子的数量和大小来改变分组的细度和粗细，以适应不同的数据特征。直方图也经常与其他统计图表一起使用，比如箱线图、散点图、概率密度估计曲线等，以便更深入地理解数据的特征和结构。

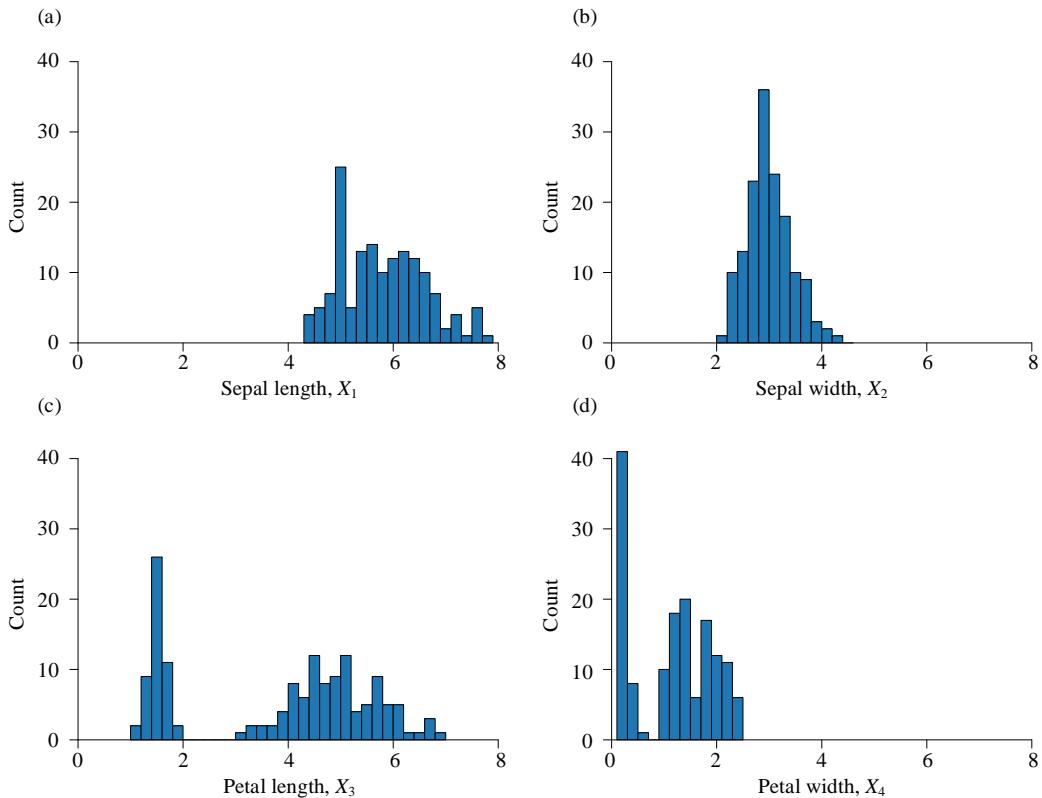


图 4. 鸢尾花四个特征数据的直方图，纵轴为频数

图 5 所示为同一个坐标系下对比鸢尾花四个特征数据直方图。图 5 (a) 纵轴为频数，图 5 (b) 纵轴为概率密度。

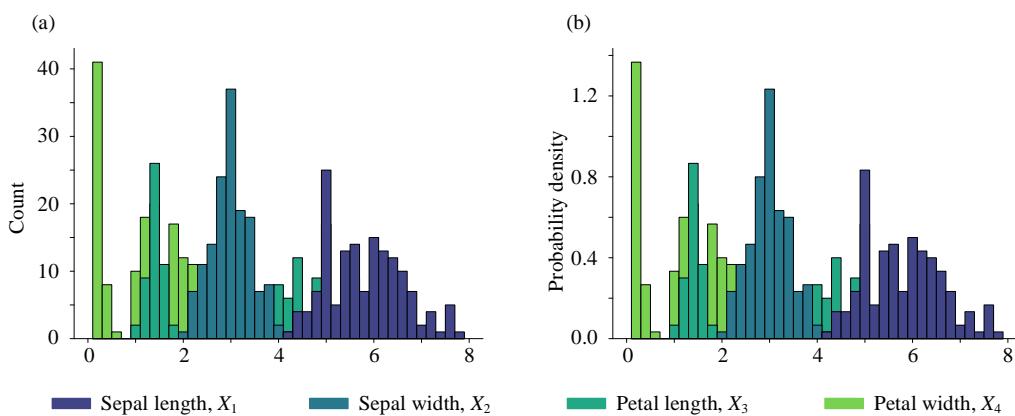


图 5. 直方图，比较频数和概率密度

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

累积频数、累积概率

图 6 对比四个鸢尾花特征样本数据的累积频数图、累积概率图。如图 6 (a) 所示，累积频数的最大值为 150，即鸢尾花数据集样本个数。如图 6 (b) 所示，累积概率的最大值为 1。

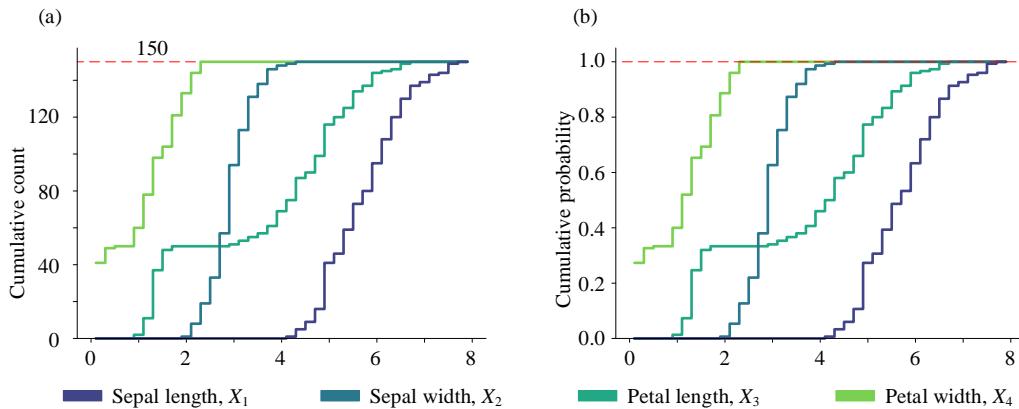


图 6. 累积频数图、累积概率图

多边形图、概率密度估计

多边形图 (polygon) 将直方图矩形顶端中点连接，得到如图 7 (a) 所示线图。

⚠ 注意，多边形图的纵轴和直方图一样有很多选择，图 7 (a) 给出的纵轴为概率密度。

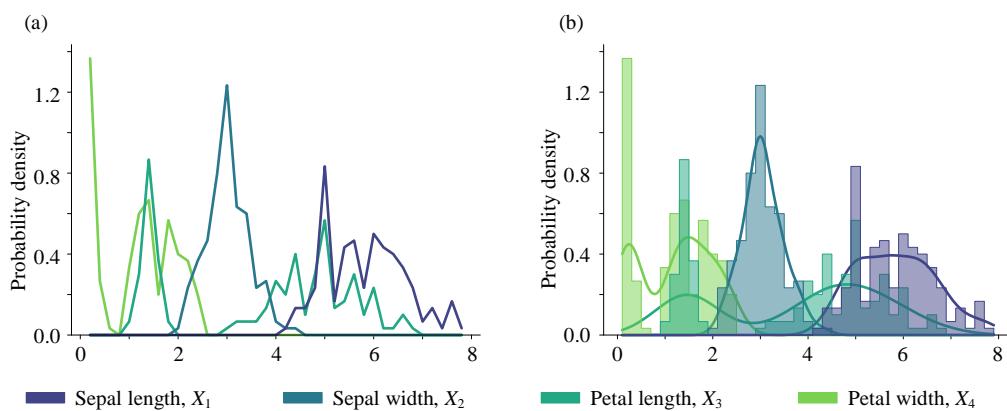


图 7. 比较多边形图和概率密度估计曲线

核密度估计 (Kernel Density Estimation, KDE) 是对直方图的扩展，如图 7 (b) 中曲线是通过核密度估计得到的概率密度函数图像。

概率密度函数描述的是随机变量在某个取值点的概率密度，是描述随机变量分布的基本函数之一。在实际问题中，往往无法直接获得概率密度函数，因此需要通过概率密度估计来估计概率密度函数。概率密度估计可以通过多种方法来实现，比如直方图法、参数法、核密度估计法、最大似然估计法等。其中，核密度估计法是最常用的方法之一，它假设数据的概率密度函数是由一些基本的核函数叠加而成，然后根据数据样本来确定核函数的带宽和数量，最终得到概率密度函数的估计值。

→ 本书第 9、10 章介绍用高斯分布完成概率密度估计，第 17 章将专门讲解概率密度估计。

山脊图

山脊图 (ridgeline plot) 是由多个重叠的概率密度线图构成，这种可视化方案形式上紧凑。图 8 所示的山脊图采用 Jopyy 绘制。

山脊图的基本思想是，将数据沿着 y 轴的方向上的一条带状区间内进行展示，使得数据的分布曲线能够清晰地显示出来，并且不会重叠和遮挡。在山脊图中，每个变量的分布曲线通常用核密度估计法或直方图法进行估计，然后按照一定的顺序进行平移和叠加。

山脊图常用于探索多个变量之间的关系和相互作用，以及发现变量的共同分布特征和异常点。它可以用于可视化各种类型的数据，比如时间序列数据、连续变量数据、分类变量数据等。

→ 本书第 20、21 章将利用山脊图可视化后验概率连续变化。

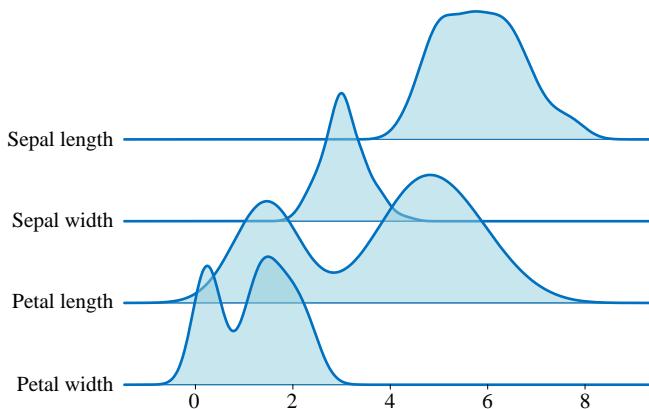


图 8. 鸢尾花数据山脊图

2.3 散点图：两特征数据分布

二维数据最基本的可视化方案是**散点图** (scatter plot)，如图 9 (a) 所示。散点图常用于展示两个变量之间的关系和相互作用。散点图将每个数据点表示为二维坐标系上的一个点，其中一个变量沿 x 轴方向表示，另一个变量沿 y 轴方向表示，每个点的位置反映了两个变量之间的数值关系。

散点图可以用于研究两个变量之间的线性关系、非线性关系或者无关系。如果两个变量之间存在线性关系，那么散点图中的点会形成一条斜率为正或负的回归直线。如果两个变量之间存在非线性关系，那么散点图中的点会形成一条回归曲线或者散布在二维坐标系的不同区域。如果两个变量之间无关系，那么散点图中的点会相对均匀地分布在二维坐标系中。

 本书第 24 章将介绍线性回归。此外，《数据有道》一册将专门讲解各种常见回归模型。

散点图常用于探索数据中的异常值、趋势和模式，并且可以发现变量之间的相互作用和关联性。

在散点图的基础上，我们可以拓展得到一系列衍生图像。比如图 9 (a) 中，我们可以看到两幅**边缘直方图** (marginal histogram)，它们分别描绘花萼长度和花萼宽度这两个特征的分布状况。图 9 (b) 增加了简单线性回归图像和边缘 KDE 概率密度曲线。

边缘概率 (marginal probability) 和**联合概率** (joint probability) 相对应。联合概率针对两个及以上随机变量的分布，边缘概率对应单个随机变量。图 9 中两幅图一方面展示两个随机变量的联合分布，同时展示每个随机变量的单独分布。大家会在本书后续经常看到类似的可视化方案。

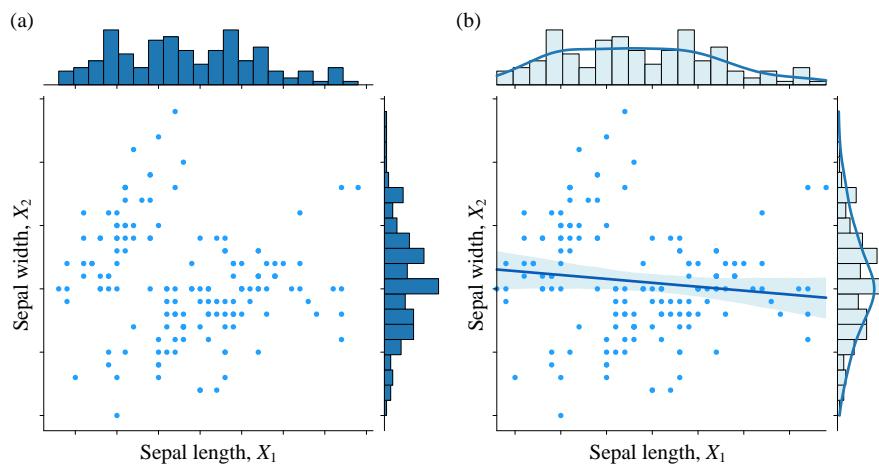


图 9. 二维数据散点图及扩展

二维概率密度

我们可以将上一节的直方图和 KDE 概率密度曲线，都拓展到二维数据。图 10 (a) 所示为二维直方图热图，热图每一个色块的颜色深浅代表该区域样本数据的频数。图 10 (b) 为二维 KDE 概率密度曲面等高线图。

图 11 (a) 在直方图热图上增加了边缘直方图，图 11 (b) 在二维联合概率密度曲面等高线图上增加了边缘概率密度曲线。

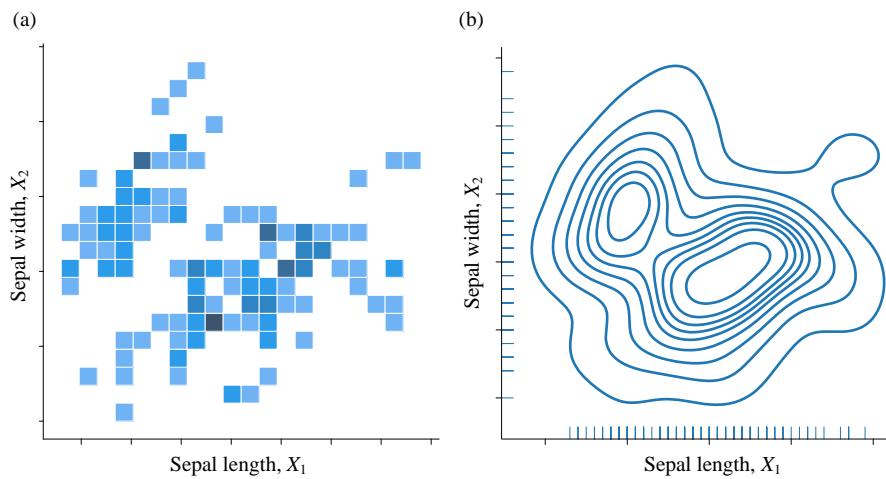


图 10. 二维数据直方图热图，二维 KDE 概率密度曲面等高线

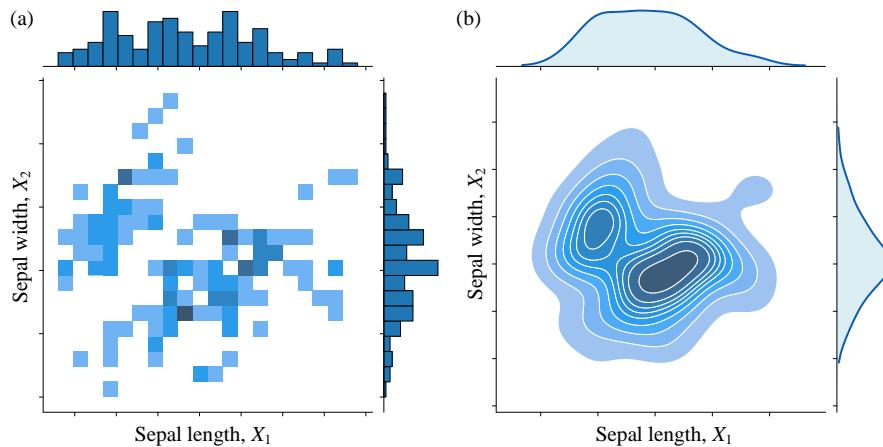


图 11. 直方图热图和概率密度曲面等高线拓展

成对特征图

本节介绍的几种二维数据统计分析可视化方案也可以拓展到多维数据，图 12 所示为鸢尾花数据成对特征分析图。鸢尾花书读者对图 12 已经完全不陌生，我们在《数学要素》、《矩阵力量》都讲过成对特征分析图。

图 12 这幅图像有 4×4 个子图，主对角线上的图像为鸢尾花单一特征数据直方图，右上角六幅子图为成对数据散点图，左下角六幅子图为概率密度曲面等高线图。

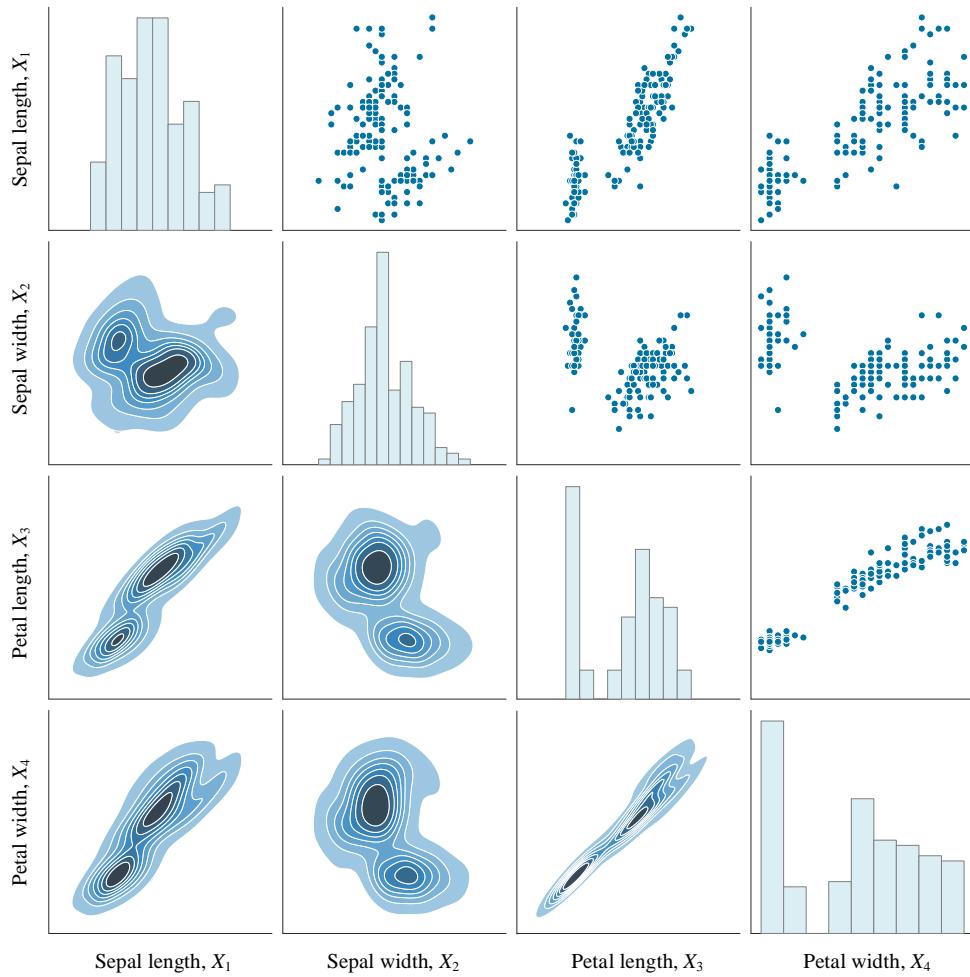


图 12. 鸢尾花数据成对特征分析图

2.4 有标签数据的统计可视化

《矩阵力量》专门区分过**有标签数据** (labeled data) 和**无标签数据** (unlabeled data)，如图 13 所示。

鸢尾花数据就是典型的有标签数据。鸢尾花数据有三个标签——**山鸢尾** (setosa)、**变色鸢尾** (versicolor) 和**维吉尼亚鸢尾** (virginica)。每一行样本点都对应特定鸢尾花分类。

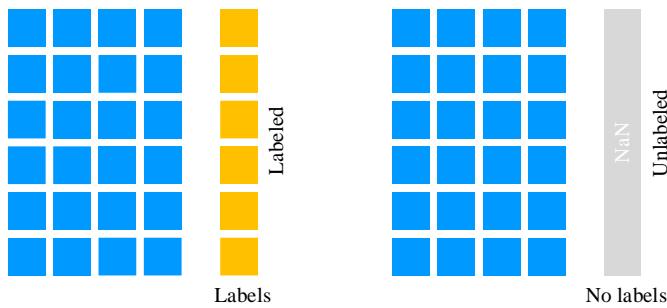


图 13. 根据有无标签分类数据

图 14 所示为含有标签分类的直方图。不同类别的鸢尾花数据采用不同颜色的直方图。图 14 的纵轴可以是频数、概率、概率密度。此外，考虑到分类标签，概率、概率密度也可以对应条件概率。举个例子，如果图 14 的纵轴对应“条件”概率密度的话，每幅子图中不同颜色的直方图面积均为 1。

条件概率中的“条件”听起来很迷惑，实际上大家在生活中经常用到。比如，高中二年 3 班男生的平均身高，“高中二年 3 班”和“男生”都是条件。不难理解，“条件”实际上就是限定讨论范围。

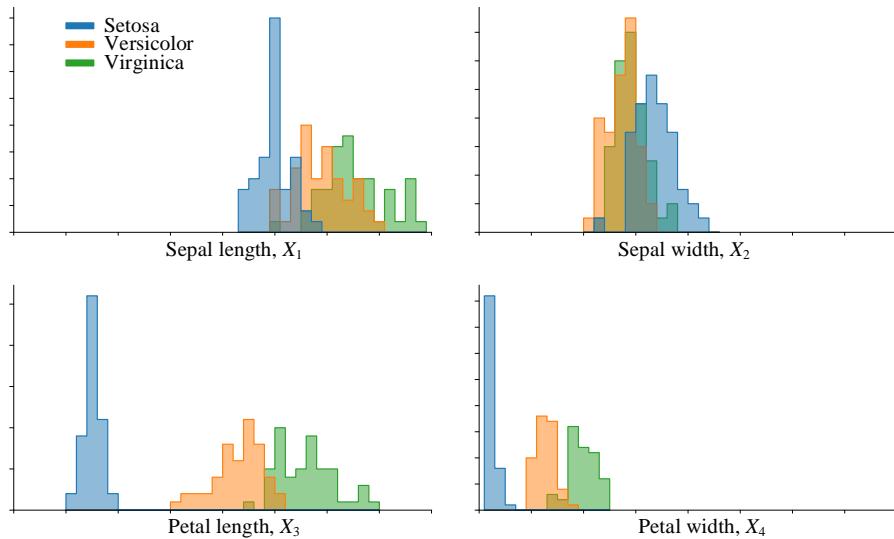


图 14. 直方图，考虑鸢尾花分类标签

图 15 所示为考虑分类的山脊图。我们也可以把这种可视化方案应用到二维数据可视化，如图 16 所示。图 17 所示为考虑标签的成对特征图。

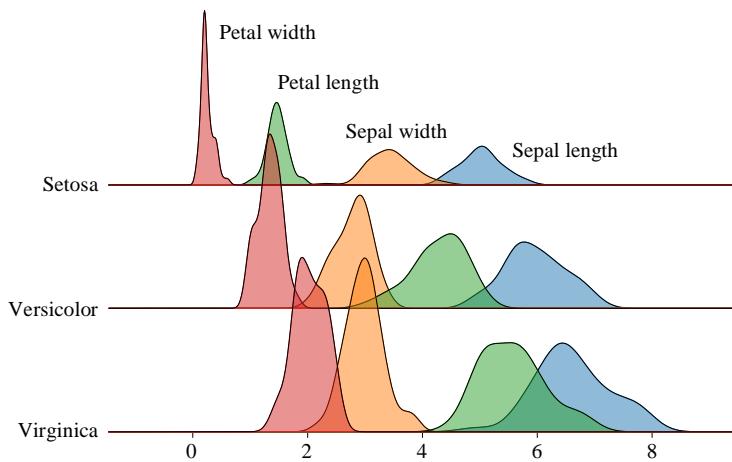


图 15. 鸢尾花山数据山脊图，特征分类

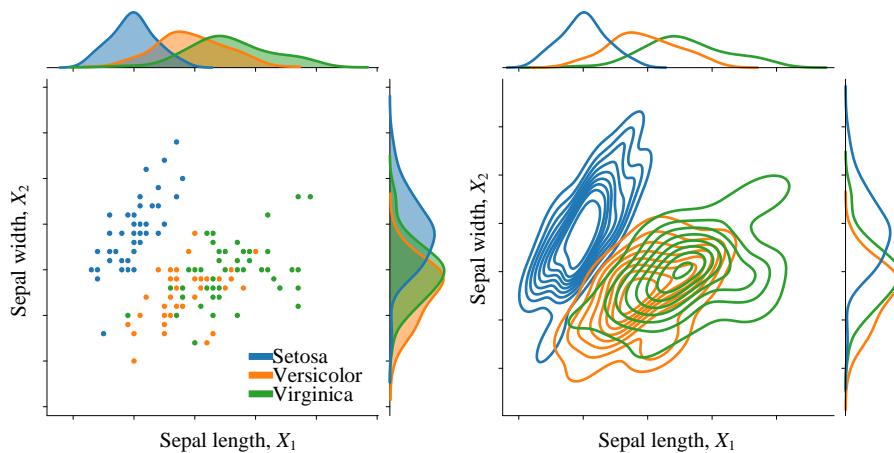


图 16. 二维数据散点图，KDE 概率密度曲面等高线，考虑鸢尾花分类标签

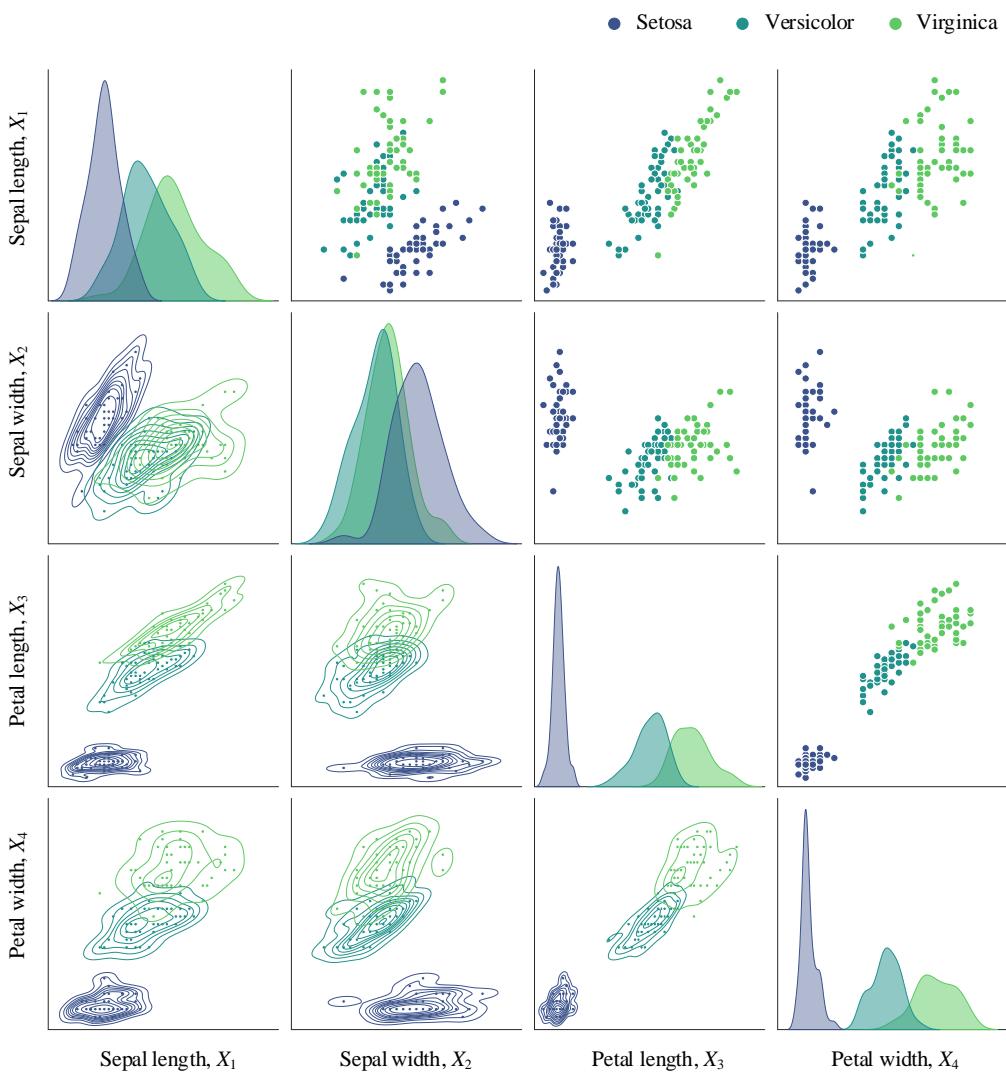


图 17. 鸢尾花数据成对特征分析图，考虑鸢尾花分类标签

平行坐标图

平行坐标图 (Parallel Coordinate Plot, PCP) 能够在二维空间中呈现出多维数据。在平行坐标图中，每条折线代表一个样本点，图中每条竖线代表一个特征。折线的形状能够反映样本的若干特征。不同折线颜色代表不同分类标签，平行坐标图还可以不同特征对分类的影响。

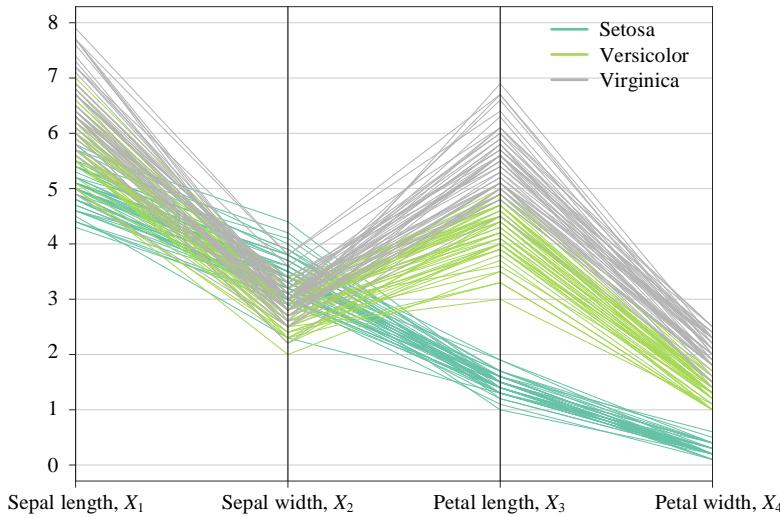


图 18. 鸢尾花数据的平行坐标图

2.5 集中度：均值、质心

本章前文通过图形可视化样本分布，本章后文介绍几种最基本的量化手段来描述样本数据。

量化样本数据集中度的最基本方法是**算数平均数** (arithmetic mean)：

$$\mu_x = \text{mean}(X) = \frac{1}{n} \left(\sum_{i=1}^n x^{(i)} \right) = \frac{x^{(1)} + x^{(2)} + x^{(3)} + \dots + x^{(n)}}{n} \quad (6)$$

如果数据是总体，算数平均数为**总体平均值** (population mean)。如果数据是样本，算数平均数是**样本平均值** (sample mean)。

注意，计算均值时，(6) 中每个样本的权重相同，都是 $1/n$ 。本书后续大家会发现，对于离散随机变量，权重由概率质量函数决定。



请大家回顾《矩阵力量》第 22 章讲过的均值的几何意义。

以鸢尾花数据集为例

鸢尾花四个量化特征——花萼长度 (sepal length) X_1 、花萼宽度 (sepal width) X_2 、花瓣长度 (petal length) X_3 和花瓣宽度 (petal width) X_4 ——均值分别为：

$$\mu_1 = 5.843, \mu_2 = 3.057, \mu_3 = 3.758, \mu_4 = 1.199 \quad (7)$$

图 4 所示为鸢尾花数据集四个特征均值在直方图位置。

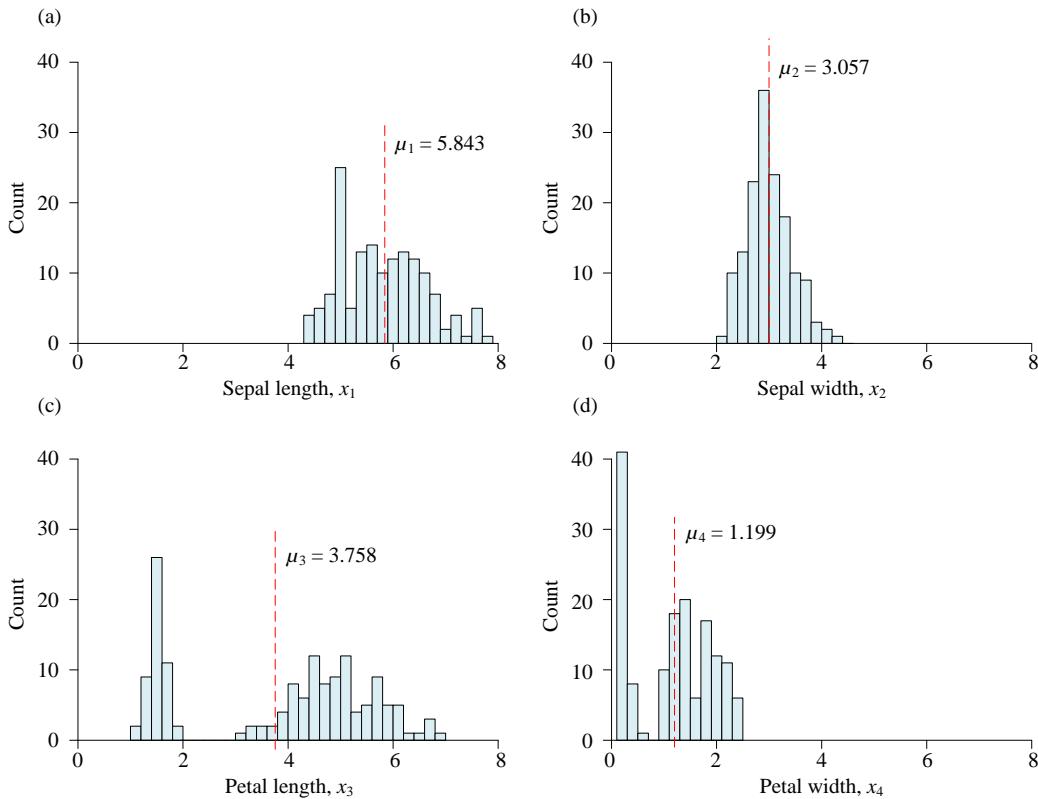


图 19. 鸢尾花四个特征数据均值在直方图位置

质心

当然，我们也可以把均值位置标注在散点图上。如图 20 所示，花萼长度、花萼宽度的均值相交于一点 \times ，这一点常被称作数据的**质心** (centroid)。也就是说，有些场合，我们可以用质心这个点代表一组样本数据。

比如，鸢尾花数据矩阵 X 质心为：

$$\mathbb{E}(X) = \boldsymbol{\mu}_X^T = \begin{bmatrix} 5.843 & 3.057 & 3.758 & 1.199 \\ \text{Sepal length, } x_1 & \text{Sepal width, } x_2 & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{bmatrix}^T \quad (8)$$

本书中， $\mathbb{E}(X)$ 一般为行向量，而 $\boldsymbol{\mu}$ 一般为列向量。此外，本书一般不从符号上区别样本均值和总体均值（期望值），除非特别说明。

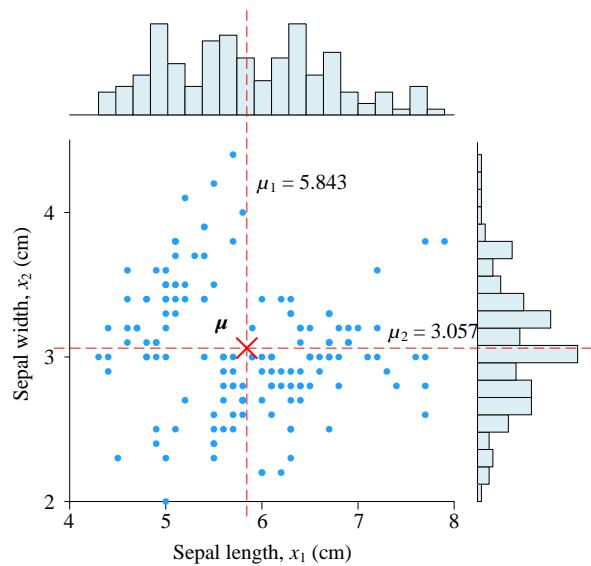


图 20. 均值在散点图的位置

考虑分类标签

分别计算鸢尾花不同分类标签 (setosa、versicolor、virginica) 花萼长度、花萼宽度平均值：

$$\begin{aligned}\mu_{1_setosa} &= 5.006, \quad \mu_{2_setosa} = 3.428 \\ \mu_{1_versicolor} &= 5.936, \quad \mu_{2_versicolor} = 2.770 \\ \mu_{1_virginica} &= 6.588, \quad \mu_{2_virginica} = 2.974\end{aligned}\tag{9}$$

图 21 所示为不同分类标签的鸢尾花样本散点，以及各自的簇质心 (cluster centroid)。

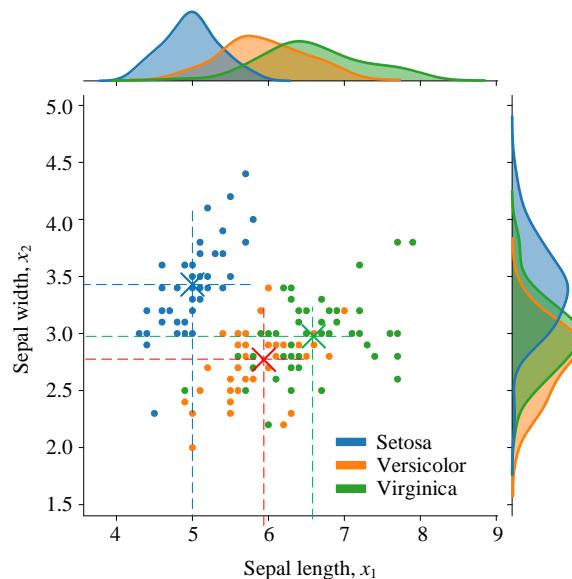


图 21. 均值在散点图的位置，考虑类别标签

中位数、众数、几何平均数

中位数 (median) 又称中值，指的是按顺序排列的一组样本数据中居于中间位置的数。如果样本数量为奇数，从小到大排列居中的样本就是中位数；如果样本有偶数个，通常取最中间的两个数值的平均数作为中位数。



本书后续将在贝叶斯推断中进一步比较均值、中位数。

众数 (mode) 是一组数中出现最频繁的数值。众数通常用于描述离散型数据，因为这些数据中每个值只能出现整数次，而众数是出现次数最多的值。对于连续型数据，比如身高、体重，由于每个数值只有极小的概率出现，因此通常不会存在一个数值出现次数最多的情况，此时可以使用**区间众数** (interval mode) 来描述数据的分布形态。

众数的计算相对简单，只需要统计每个数值出现的次数，然后找到出现次数最多的数值即可。众数的缺点是可能存在多个众数或者无众数的情况，而且受极端值的影响较大。

几何平均数 (geometric mean) 的定义如下：

$$\left(\prod_{i=1}^n x^{(i)} \right)^{\frac{1}{n}} = \sqrt[n]{x^{(1)} \cdot x^{(2)} \cdot x^{(3)} \cdots x^{(n)}}$$

⚠ 注意，几何平均数只适合正数。

2.6 分散度：极差、方差、标准差

本节介绍度量分散度的常见统计量。

极差

极差 (range)，又称全距，是指样本最大值与最小值之间的差距：

$$\text{range}(X) = \max(X) - \min(X) \quad (10)$$

极差是度量分散度最简单的指标。图 22 所示为最大值、最小值、极差、均值之间关系。注意，极差很容易受到离群值影响。

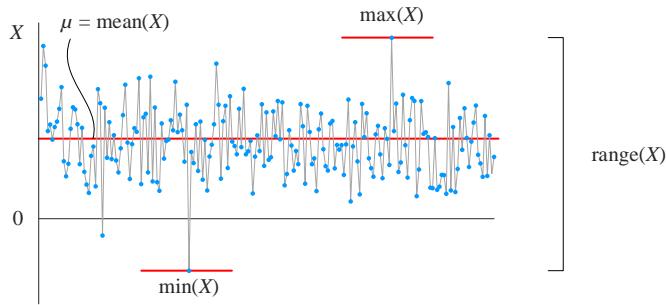


图 22. 最大值、最小值、极差、均值的关系

方差

方差 (variance) 衡量随机变量或样本数据离散程度。方差越大，数据的分布就越分散；方差越小，数据的分布就越集中。样本的方差为：

$$\text{var}(X) = \sigma_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)^2 \quad (11)$$

简单来说，方差是各观察值与数据集平均值的差的平方的平均值。

方差的单位是样本单位的平方，比如鸢尾花数据方差单位为 cm²。

⚠ 请大家注意，本书中样本方差、总体方差符号上完全一致，不做特别区分。

→ 此外，请大家回顾《矩阵力量》第 22 章讲过的方差的几何意义。

标准差

样本的**标准差** (standard deviation) 为样本方差的平方根：

$$\sigma_X = \text{std}(X) = \sqrt{\text{var}(X)} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_X)^2} \quad (12)$$

同样，标准差越大，数据的分布就越分散；标准差越小，数据的分布就越集中。鸢尾花样本数据四个量化特征的标准差分别为：

$$\sigma_1 = 0.825, \sigma_2 = 0.434, \sigma_3 = 1.759, \sigma_4 = 0.759 \quad (13)$$

⚠ 注意，标准差和原始数据单位一致。比如，鸢尾花四个特征的量化数据单位均为厘米 (cm)。

图 23 上，我们把 $\mu \pm \sigma$ 、 $\mu \pm 2\sigma$ 对应的位置也画在直方图上。

→ 68-95-99.7 法则和 $\mu \pm \sigma$ 、 $\mu \pm 2\sigma$ 、 $\mu \pm 3\sigma$ 有关，本书第 9 章将介绍 68-95-99.7 法则。

其实，大家在生活中经常用到“均值”、“标准差”这两个概念，只不过大家没有注意到而已。举个例子，想要提高考试成绩，大家平时练习时会尽量提高平均分，并减小各种因素让自己发挥稳定。这就是在增大均值，减小标准差（波动）。

再举个例子，一个教练在选择哪个选手上场的时候，也会看“均值”、“标准差”。“均值”代表一个选手的绝对实力，“标准差”则代表选手成绩的波动幅度。

教练求稳的时候，会派出均值相对高、标准差（波动）小的选手。在大比分落后情况下，教练可能会派出临场发挥型选手。发挥型选手成绩均值可能不是最高，但是有机会“冲一冲”。

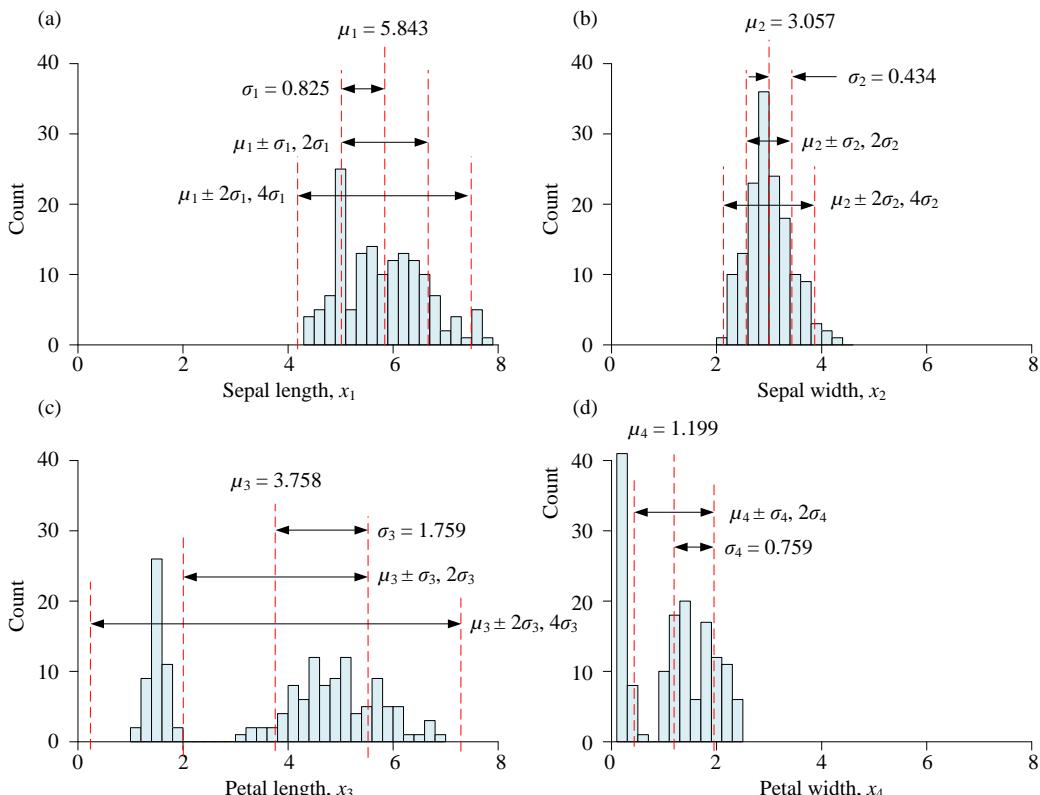


图 23. 鸢尾花四个特征数据均值、标准差所在位置在直方图位置

2.7 分位：四分位、百分位等

分位数 (quantile)，亦称分位点，是指将一个随机变量的概率分布范围分为几个等份的数值点。常用的分位数有**二分位点** (2-quantile, median)、**四分位点** (4-quantiles, quartiles)、**五分位点** (5-quantiles, quintiles)、**八分位点** (8-quantiles, octiles)、**十分位点** (10-quantiles, deciles)、**二十分位点** (20-quantiles, vigintiles)、**百分位点** (100-quantiles, percentile) 等。

实践中，四分位和百分位最常用。以百分位为例，把一组从小到大排列的样本数据分为 100 等份后，每一个分点就是一个百分位数。

同理，将所有样本数据从小到大排列，四分位数对应三个分割位置（25%、50%、75%）。这三个分割位置将样本平分为四等份。50% 分位数对应中位数。图 24 所示为将鸢尾花不同特征的四分位画在直方图上。

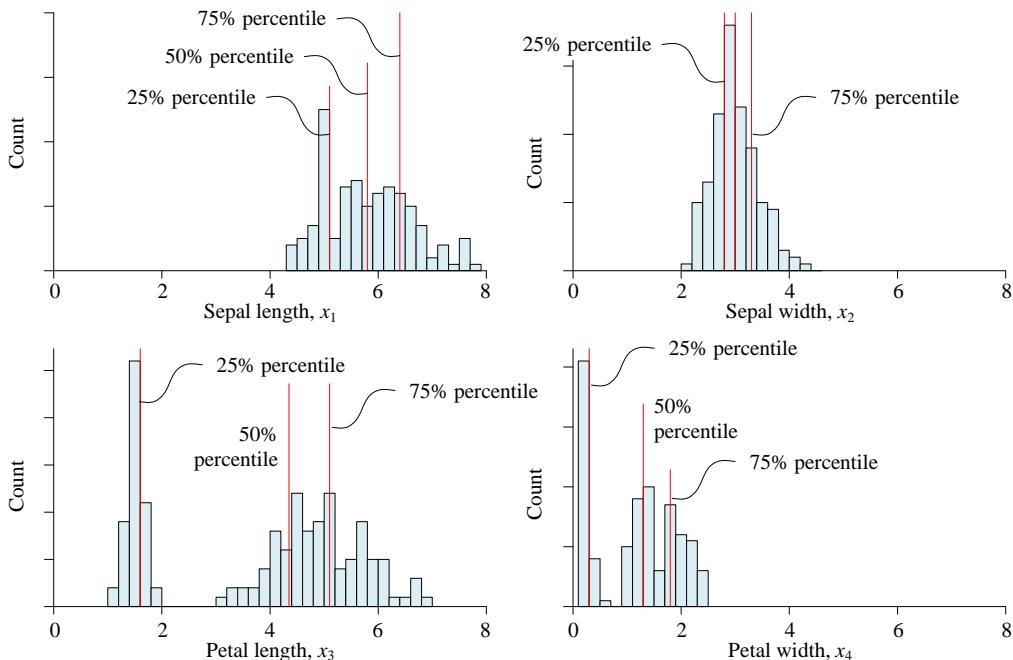


图 24. 鸢尾花数据直方图，以及 25%、50% 和 75% 百分位

图 25 所示为鸢尾花四个特征数据 1%、50%、99% 三个百分位位置，1%、99% 可以用来描述样本分布的“左尾”、“右尾”。

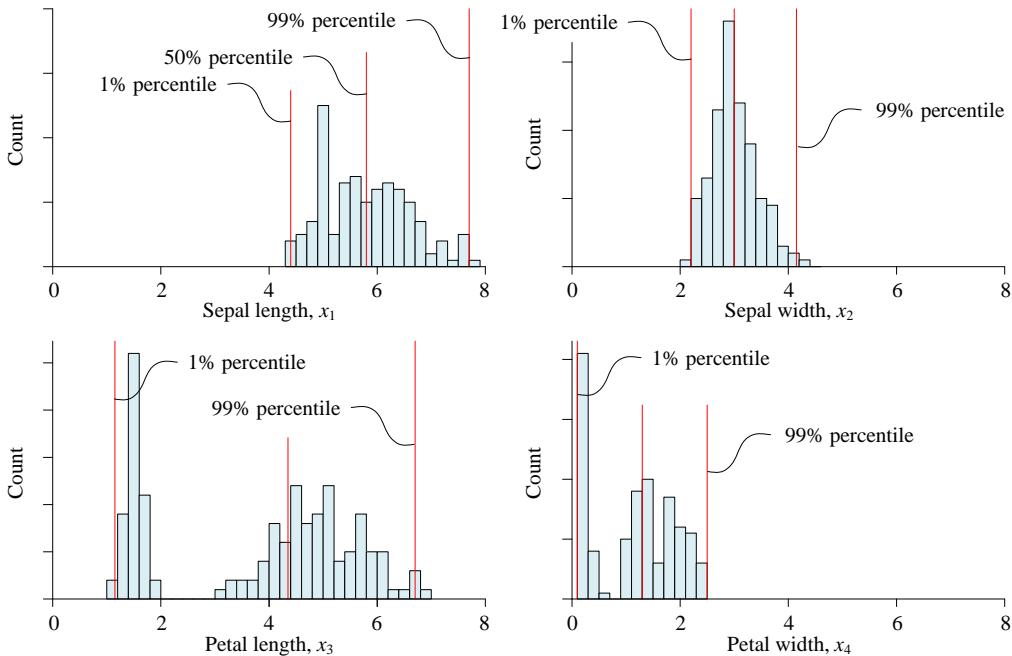


图 25. 鸢尾花数据直方图，以及 1% 和 99% 百分位

对于 Pandas 数据帧 df, df.describe() 默认输出数据的样本总数、均值、标准差、最小值、25% 分位、50% 分位(中位数)、75% 分位。图 26 所示鸢尾花数据帧的总结，其中还给出 1% 百分位、99% 百分位。

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
1%	4.400000	2.200000	1.149000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
99%	7.700000	4.151000	6.700000	2.500000
max	7.900000	4.400000	6.900000	2.500000

图 26. 鸢尾花数据帧统计总结

2.8 箱型图：小提琴图、分布散点图

图 27 所示为 **箱型图** (box plot) 原理。箱型图利用第一 ($25\%, Q_1$)、第二 ($50\%, Q_2$) 和第三 ($75\%, Q_3$) 四分位数展示数据分散情况。 Q_1 也叫下四分位， Q_2 也叫中位数， Q_3 也称上四分位。

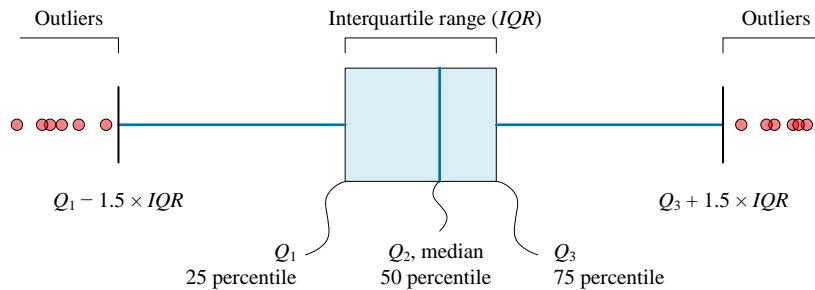


图 27. 箱型图原理

箱型图的**四分位间距** (interquartile range) 定义为：

$$IQR = Q_3 - Q_1 \quad (14)$$

箱型图也常用来分析样本中可能存在的离群点，图 27 中两侧的红点。 $Q_3 + 1.5 \times IQR$ 也称上界， $Q_1 - 1.5 \times IQR$ 叫下界。而在 $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$ 之外的样本数据则被视作离群点。

数据分析中，四分位间距 IQR 也常常用来度量样本数据的分散程度。相比标准差，四分位间距 IQR 不受厚尾影响，受离群值影响小得多。

图 28 所示为鸢尾花数据四个特征上的箱型图。

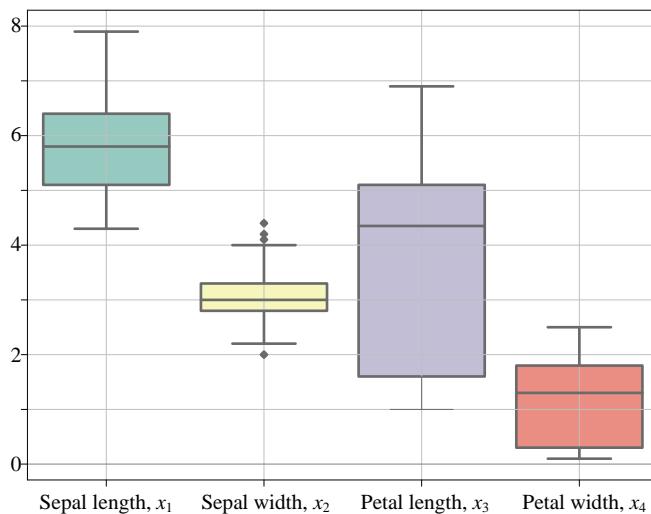


图 28. 鸢尾花数据箱型图

箱型图的变体

箱型图还有很多的“变体”。比如图 29 所示的小提琴图，图 30 所示的分布散点图。图 31 所示为箱型图叠加分布散点图。图 32 所示为考虑标签的箱型图。

箱型图的优点是简单易懂，可以同时展示数据的中心趋势、离散程度和离群值等信息。因此，箱型图经常被用来比较多组数据的分布情况，或者发现异常值。

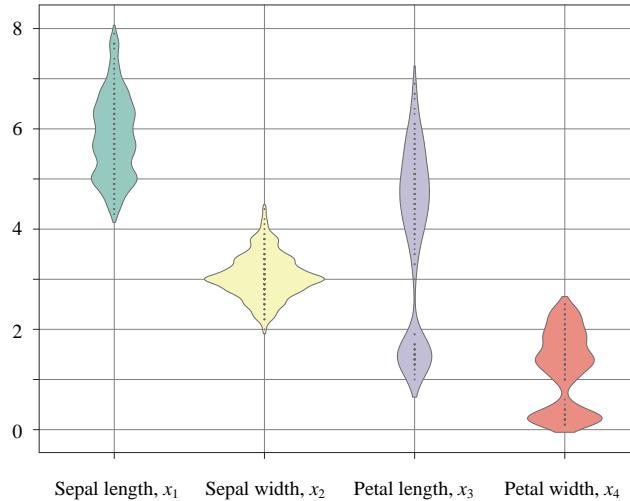


图 29. 鸢尾花数据小提琴图

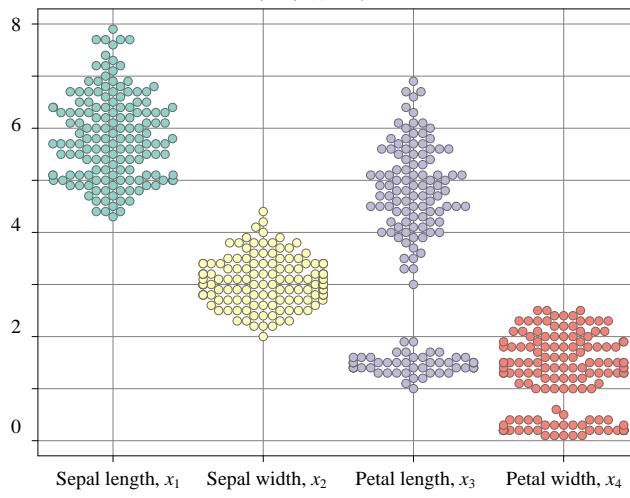


图 30. 分布散点图 (stripplot)

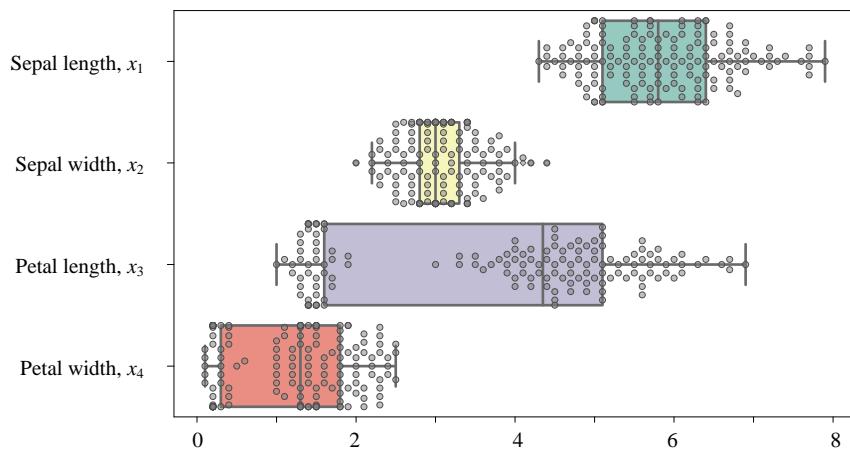


图 31. 鸢尾花箱型图，叠加分布散点图 swarmplot

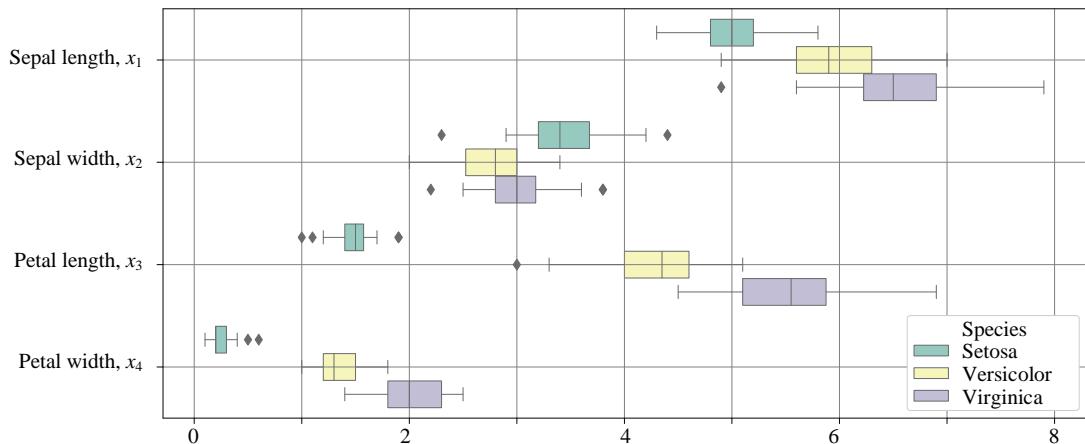


图 32. 鸢尾花箱型图，考虑分类标签

2.9 中心距：均值、方差、偏度、峰度

统计学中的**矩**(moment)，又称为**中心矩**(central moment)，是对变量分布和形态特点进行度量的一组量，其概念借鉴物理学中的“矩”。在物理学中，矩是描述物理性状特点的物理量。

零阶矩表示随机变量的总概率，也就是 1。具体而言，常用的中心矩为一至四阶矩，分别表示数据分布的位置、分散度、偏斜程度和峰度程度。

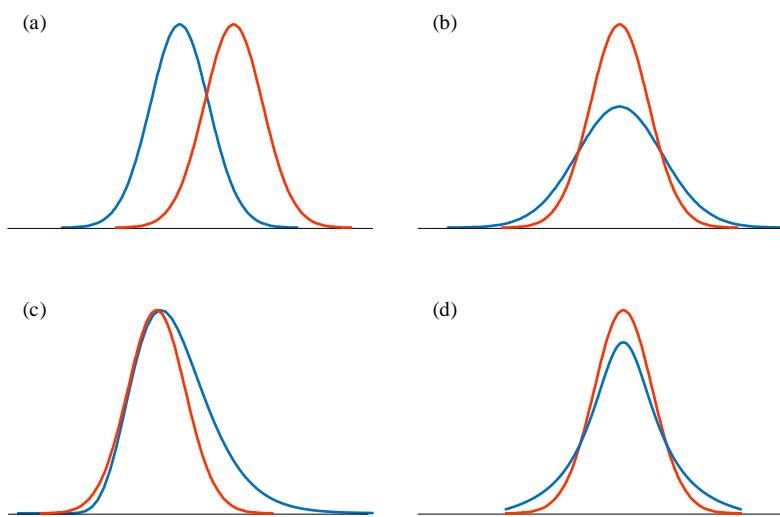


图 33. 期望(一阶矩)、方差(二阶矩)、偏斜度(三阶矩)、峰度(四阶矩)

一阶矩、二阶矩

一阶矩为均值，即**期望**(expectation)，用来描述分布中心位置，如图 33 (a) 所示。前文提过，均值的量纲(单位)和原始数据相同。

⚠ 注意，量纲和单位虽然混用，但是两者还是有区别。从量纲的角度来看，m、cm、mm 都是长度度量单位，含义相同。但是，m、cm、mm 的单位不同，它们之间存在一定换算关系。

二阶矩为**方差**(variance)，描述分布分散情况，如图 33 (b) 所示。方差的量纲为原始数据量纲的平方。一元高斯分布的参数仅为均值和方差。

均值和方差都相同不能说明分布相同。换个角度，真实的样本数据分布不可能仅仅用均值和方差来刻画，有时还需要偏度(三阶矩)和峰度(四阶矩)。

三阶矩

三阶矩为**偏度**(skewness) S 。如图 33 (c) 所示，偏度描述分布的左右倾斜程度：

$$S = \text{skewness} = \frac{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^3}{\left(\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^2 \right)^{\frac{3}{2}}} \quad (15)$$

与期望和标准差不同，偏度没有单位，是无量纲量。偏度的绝对值越大，表明样本数据分布的偏斜程度越大。

对于完全对称的单峰分布，平均数、中位数、众数，处在同一位置，图 34 (a) 所示。这种分布的偏度为零。如果样本数服从一元高斯分布，则偏度为 0，即均值 = 中位数 = 众数。

正偏 (positive skew, positively skewed), 又称**右偏** (right-skewed, right-tailed, skewed to the right)。如图 34 (b) 所示, 正偏分布的右侧尾部更长, 分布的主体集中在图像的左侧。正偏(右偏)时, 均值 > 中位数 > 众数。

大家可以这样理解平均数、中位数、众数这三个数值的关系。如果在样本中引入少数几个特别大的离群值的话, 均值肯定增大(向右移动), 中位数微微受到影响(样本数量增加), 但是众数(出现次数最多)不变。

负偏 (negative skew, negatively skewed), 又称**左偏** (left-skewed, left-tailed, skewed to the left), 如图 34 (c) 所示, 特点是分布的左侧尾部更长, 分布的主体集中在右侧。负偏(左偏)时, 众数 > 中位数 > 均值。

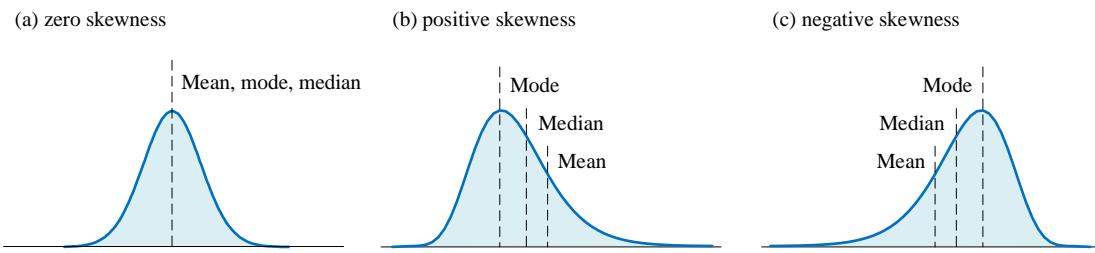


图 34. 无偏、正偏和负偏

⚠ 值得注意的是, 偏度为零不一定意味着分布对称。如图 35 所示, 这个离散分布的偏度计算出来为 0, 但是很明显这个分布不对称。

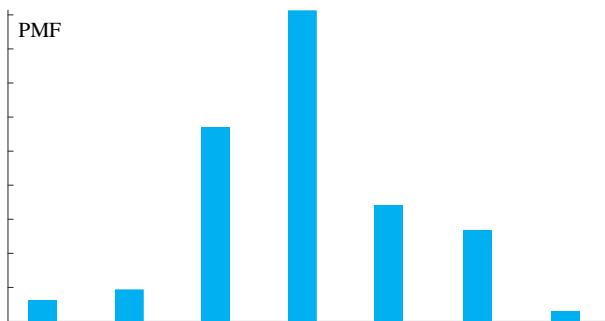


图 35. 偏度为 0, 但是不对称的分布

四阶矩

四阶矩表示**峰度** (kurtosis) K 。图 33 (d) 所示, 峰度描述分布与正态分布相比的陡峭或扁平程度:

$$K = \text{kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^2 \right)^2} \quad (16)$$

和偏度一样，峰度也没有单位，是无量纲量。

⚠ 注意，用 (16) 计算的话，正态分布的峰度为 3。

图 36 展示两种峰态：**高峰态** (leptokurtic)、**低峰态** (platykurtic)。高峰度的峰度值大于 3。如图 36 (a) 所示，和正态分布相比，高峰态分布有明显的尖峰，两侧尾端有**肥尾** (fat tail)。

图 36 (b) 展示的是低峰态。相比正态分布，低峰态明显稍扁。

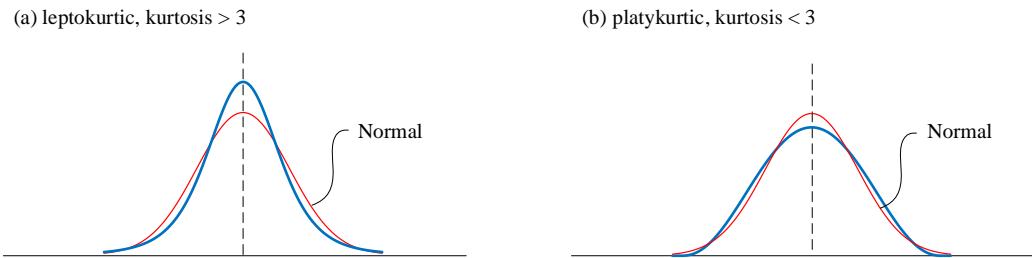


图 36. 高峰态和低峰态

实践中，一般采用**超值峰度** (excess kurtosis)，即 (16) 减去 3：

$$\text{Excess kurtosis} = \frac{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_x)^2 \right)^2} - 3 \quad (17)$$

“减去 3”是为了让正态分布的峰度为 0，方便其他分布和正态分布比较。

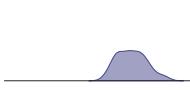
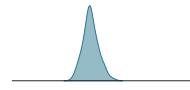
表 1 总结鸢尾花数据的四阶矩。花萼长度、花萼宽度上，样本数据都存在正偏。花萼长度分布存在低峰态，花萼宽度上出现高峰态。

对比表中样本数据分布和四阶矩的具体值，不难发现即便使用四阶矩也未必能够准确描述真实分布。比如，花瓣长度、花瓣宽度上，样本数据分布存在明显的双峰态。



本书第 9 章讲解 QQ 图时，还会提到不同的分布类型。

表 1. 鸢尾花四阶矩

				
花萼长度				
均值 (cm)	5.843	3.057	3.758	1.199
标准差 (cm)	0.825	0.434	1.759	0.759
偏度	0.314	0.318	-0.274	-0.102
超值峰度	-0.552	0.228	-1.402	-1.340

2.10 多元随机变量关系：协方差矩阵、相关性系数矩阵

协方差 (covariance) 是用来度量两个变量之间的线性关系强度和方向的统计量。当两个变量的协方差为正时，说明它们的变化趋势同向，即当一个变量增加时，另一个变量也倾向于增加；当协方差为负时，说明它们的变化趋势是相反的，即当一个变量增加时，另一个变量倾向于减少。协方差为 0 时，则表明两个变量之间没有线性关系。



鸢尾花书《数学要素》第 21 章曾图解协方差，建议大家回顾。

对于样本数据，随机变量 X 和 Y 的协方差为：

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \mu_x)(y^{(i)} - \mu_y) \quad (18)$$

线性相关性系数 (linear correlation coefficient)，也叫**皮尔逊相关系数** (Pearson correlation coefficient)，是一种用来度量两个变量之间线性相关程度的统计量。它的取值范围在 -1 到 1 之间，数值越接近 -1 或 1，表示两个变量之间的线性关系越强；数值接近 0，则表示两个变量之间没有线性关系。

对于样本数据，随机变量 X 和 Y 的线性相关性系数为：

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \quad (19)$$

“鸢尾花书”读者对**协方差矩阵** (covariance matrix)、**相关性系数矩阵** (correlation matrix) 应该非常熟悉。协方差矩阵和相关性系数矩阵都是描述多维随机变量之间关系的矩阵。



建议大家回顾《矩阵力量》中 Cholesky 分解和特征值分解协方差矩阵会产生怎样的结果。此外，也请大家回顾协方差矩阵和格拉姆矩阵的关系。

以鸢尾花四个特征为例，它的协方差矩阵为 4×4 矩阵：

$$\Sigma = \begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \text{cov}(X_1, X_3) & \text{cov}(X_1, X_4) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \text{cov}(X_2, X_3) & \text{cov}(X_2, X_4) \\ \text{cov}(X_3, X_1) & \text{cov}(X_3, X_2) & \text{cov}(X_3, X_3) & \text{cov}(X_3, X_4) \\ \text{cov}(X_4, X_1) & \text{cov}(X_4, X_2) & \text{cov}(X_4, X_3) & \text{cov}(X_4, X_4) \end{bmatrix} \quad (20)$$

其相关性系数矩阵为 4×4 ：

$$P = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \rho_{1,4} \\ \rho_{2,1} & 1 & \rho_{2,3} & \rho_{2,4} \\ \rho_{3,1} & \rho_{3,2} & 1 & \rho_{3,4} \\ \rho_{4,1} & \rho_{4,2} & \rho_{4,3} & 1 \end{bmatrix} \quad (21)$$

图 37 所示为协方差矩阵和相关性系数矩阵热图。



本书第 13 章将专门讲解协方差矩阵。

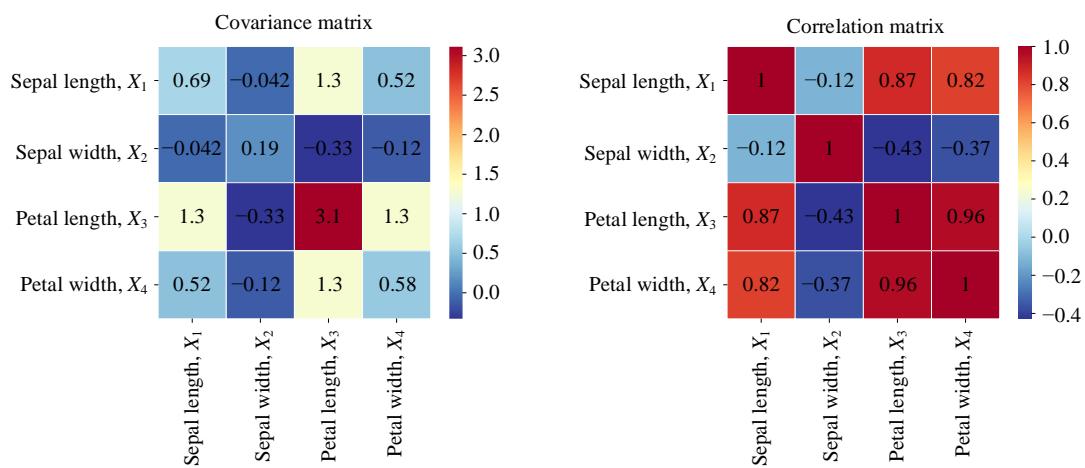
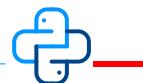


图 37. 协方差矩阵、相关性系数矩阵热图



代码文件 Bk5_Ch02_01.py 绘制本章几乎所有图像。



描述、推断是统计的两个重要板块。本章介绍了常见的统计描述工具。统计分析中，可视化和量化分析都很重要。本章介绍的重要的统计可视化工具有直方图、散点图、箱型图、热图等

等。此外，也需要大家数量掌握样本数据的均值、方差、标准差、协方差、协方差矩阵、相关系数矩阵等等。

统计描述、统计推断之间的桥梁正是概率。从下一章开始，我们正式进入概率板块学习。

3

Classical Probability

古典概率模型

归根结底，概率就是量化的生活常识



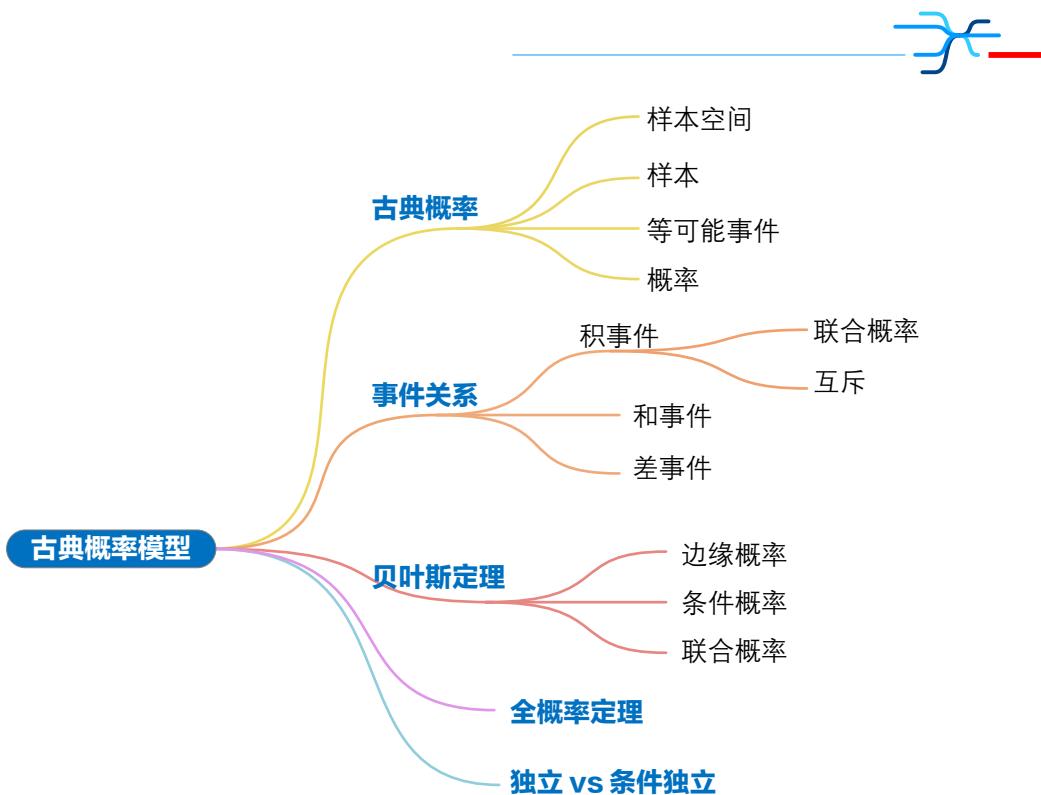
真是耐人寻味，一门以赌博为起点的学科本应该是人类知识体系中最重要研究对象。

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.

——皮埃尔-西蒙·拉普拉斯 (Pierre-Simon Laplace) | 法国著名天文学家和数学家 | 1749 ~ 1827



- ◀ numpy.array() 构造一维序列，严格来说不是行向量
- ◀ numpy.cumsum() 计算累计求和
- ◀ numpy.linspace() 在指定的间隔内，返回固定步长的数据
- ◀ numpy.random.gauss() 产生服从正态分布的随机数
- ◀ numpy.random.randint() 产生随机整数
- ◀ numpy.random.seed() 确定随机数种子
- ◀ numpy.random.shuffle() 将序列的所有元素重新随机排序
- ◀ numpy.random.uniform() 产生服从均匀分布的随机数



3.1 无处不在的概率

自然界的随机无处不在，没有两朵完全一样的鸢尾，没有两片完全一样的雪花，也没有两个完全一样的人生轨迹。鸢尾花书《数学要素》曾提过，在微观、少量、短期尺度上，我们看到的更多的是不确定、不可预测、随机；但是，站在宏观、大量、更长的时间尺度上，我们可以发现确定、模式、规律。

而概率则试图量化随机事件发生的可能性。概率的研究和应用深刻影响着人类科学发展进程，本节介绍孟德尔和道尔顿两个例子。

孟德尔的豌豆试验

孟德尔 (Gregor Mendel, 1822 ~ 1884) 之前，生物遗传机制主要是基于猜测，而不是试验。

在修道院蔬菜园里，孟德尔对不同豌豆品种进行了大量异花授粉试验。比如，孟德尔把纯种圆粒豌豆 和纯种皱粒豌豆 杂交，他发现培育得到的子代豌豆都是圆粒 ，如图 1 所示。

实际情况是，决定皱粒 的基因没有被呈现出来，因为决定皱粒 的基因相对于圆粒 基因来讲是隐性。

如图 1 所示，当第一代杂交圆粒豌豆 自花传粉或者彼此交叉传粉，它们的后代籽粒显示出 3:1 的固定比例，即 3/4 的圆粒 和 1/4 的皱粒 。

从精确的 3:1 的比例来看，孟德尔不仅仅推断出基因中离散遗传单位的存在，而且意识到这些离散的遗传单位在豌豆中成对出现，并且在形成配子的过程中分离。3:1 的比例背后的数学原理就是本章要介绍的古典概率模型。

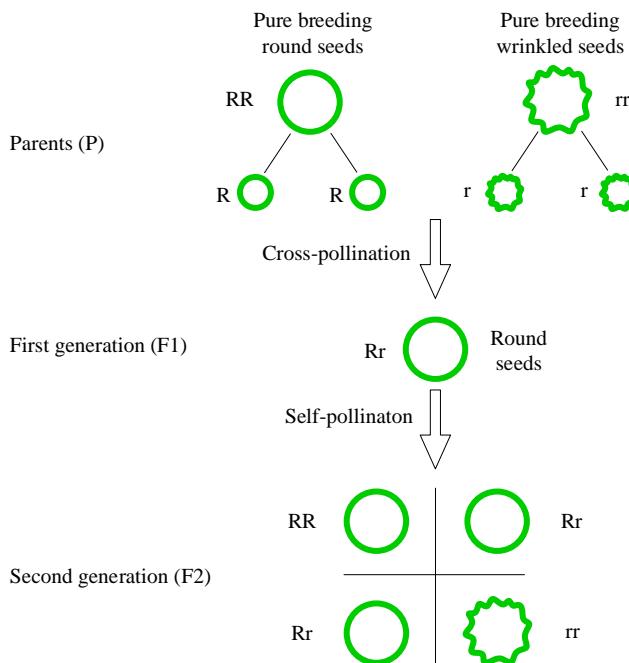


图 1. 孟德尔的豌豆试验

道尔顿发现红绿色盲

18世纪英国著名的化学家道尔顿 (John Dalton, 1766 ~ 1844) 偶然发现红绿色盲。道尔顿给母亲选了一双“棕灰色”的袜子作为圣诞礼物。但是，母亲对袜子的颜色不是很满意，她觉得“樱桃红”过于艳丽。

道尔顿十分疑惑，他问了家里的亲戚，发现只有弟弟和自己认为袜子是“棕灰色”。道尔顿意识到红绿色盲必然通过某种方式遗传。

现代人已经研究清楚，红绿色盲的遗传方式是 X 连锁隐性遗传。男性 ♂ 仅有一条 X 染色体，因此只需一个色盲基因就表现出色盲。

女性 ♀ 有两条 X 染色体，因此需有一对色盲等位基因，才会表现异常。而只有一个致病基因的女性 ♀ 只是红绿色盲基因的携带者，个体表现正常。

下面，我们从概率的角度分几种情况来思考红绿色盲的遗传规律。

情况 A

一个女性 ♀ 红绿色盲患者和一个正常男性 ♂ 生育。后代中，儿子 ♂ 都是红绿色盲；女儿 ♀ 虽表现正常，但从母亲 ♀ 获得一个红绿色盲基因，因此女儿 ♀ 都是红绿色盲基因的携带者。

不考虑性别的话，后代中发病可能性为 50%。这个可能性就是概率 (probability)。它和生男、生女的概率一致。

给定后代为男性♂，发病比例为 100%。给定后代为女性♀，发病比例为 0%，但是携带红绿色盲基因的比例为 100%。反过来，给定后代发病这个条件，可以判定后代 100% 为男性♂。这就是本章后文要介绍的**条件概率** (conditional probability)。

条件概率的概念在概率论和统计学中非常重要，它允许我们在一些已知信息的情况下对事件的发生概率进行更精确的估计和预测。例如，在医学诊断中，医生可以根据病人的症状和体征，计算出某种疾病在不同条件下的发病率，从而帮助判断病人是否患有这种疾病。

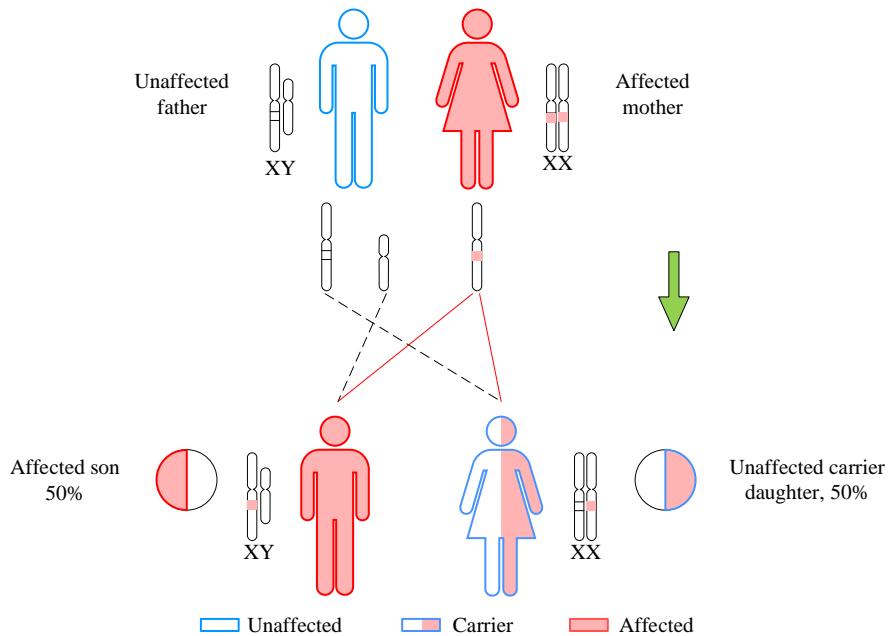


图 2. 红绿色盲基因遗传机制，情况 A

情况 B

一个女性♀红绿色盲基因携带者和一个正常男性♂生育。后代中，整体考虑，后代患病的概率为 25%。

其中，儿子♂中，50% 概率为正常，50% 概率为红绿色盲。女儿都不是色盲，但有 50% 概率是色盲基因的携带者。这些数值也都是条件概率。

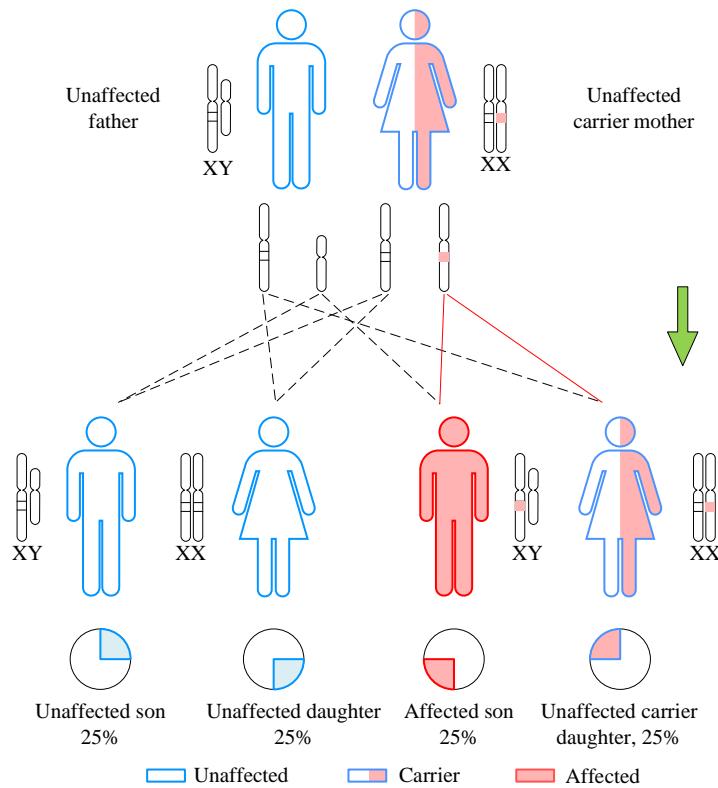


图 3. 红绿色盲基因遗传机制, 情况 B

情况 C

一个女性 ♀ 红绿色盲基因的携带者和一个男性 ♂ 红绿色盲患者生育。整体考虑来看，不分男女的话，后代发病的概率为 50%。

其中，儿子 ♂ 50% 概率正常，50% 的概率为红绿色盲。女儿 ♀ 有 50% 概率为红绿色盲，50% 概率是色盲基因的携带者。

换一个条件，如果已知后代为红绿色盲患者，后代 50% 概率为男性 ♂，50% 概率为女性 ♀。

除了以上三种情况，请大家思考还有哪些组合情况并计算后代患病概率。

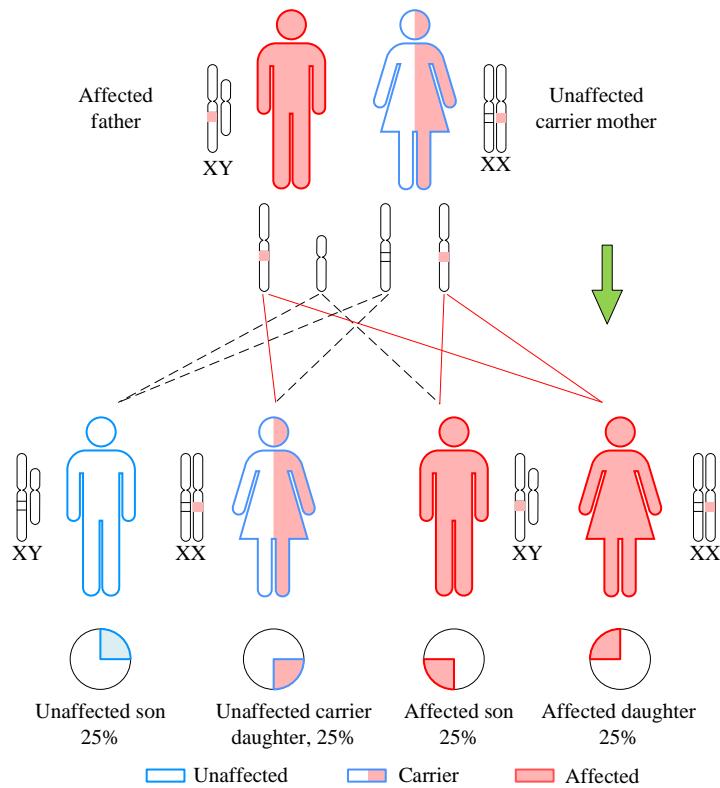


图 4. 红绿色盲基因遗传机制, 情况 C

建议大家学完本章所有内容之后，回头再琢磨孟德尔和道尔顿这两个例子。

3.2 古典概率：离散均匀概率律

概率模型是对不确定现象的数学描述。本章的核心是古典概型。古典概型，也叫等概率模型 (equiprobability)，是最经典的一种概率模型。古典模型中基本事件为有限个，并且每个基本事件为等可能。古典概型广泛应用集合运算，本节一边讲解概率论，一边回顾集合运算。



《数学要素》第 4 章介绍过集合相关概念，建议大家回顾。

给定一个随机试验，所有的结果构成的集合为**样本空间** (sample space) Ω 。样本空间 Ω 中的每一个元素为一个**样本** (sample)。不同的随机试验有各自的样本空间。样本空间作为集合，也可以划分成不同**子集** (subset)。

概率

整个样本空间 Ω 的概率为 1，即：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>
本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\Pr(\Omega) = 1$$



(1)

样本空间概率为 1，从这个视角来看，本书后续内容似乎都围绕着如何将 1“切片、切块”、“切丝、切条”。

⚠ 注意，本书表达概率的符号 \Pr 为正体。再次请大家注意，不同试验的样本空间 Ω 不同。

给定样本空间 Ω 的一个事件 (event) A ， $\Pr(A)$ 为事件 A 发生的概率 (the probability of event A occurring 或 probability of A)。 $\Pr(A)$ 满足：

$$\underbrace{\Pr}_{\text{Probability}} \left(\begin{array}{c} \text{Event} \\ A \end{array} \right) \geq 0$$



(2)

大家看到任何概率值时一定要问一嘴，它的样本空间是什么。

空集 \emptyset 不包含任何样本点，也称作**不可能事件** (impossible event)，因此对应的概率为 0：

$$\Pr(\emptyset) = 0$$

(3)

等可能

设样本空间 Ω 由 n 个**等可能事件** (equally likely events 或 events with equal probability) 构成，事件 A 的概率为：

$$\Pr(A) = \frac{n_A}{n}$$



(4)

其中， n_A 为含于事件 A 的试验结果数量。

等可能事件是指在某一试验中，每个可能的结果发生的概率相等的事件。简单来说，就是每个结果发生的可能性是一样的。例如，对于一枚硬币的抛掷，假设正面和反面的出现概率是相等的，因此正面和反面出现是等可能事件。同样地，掷一个六面骰子，假设每个面出现的概率都是相等的，因此每个面的出现也是等可能事件。

以鸢尾花数据为例

举个例子，从 150 (n) 个鸢尾花数据中取一个样本点，任何一个样本被取到的概率为 $1/150$ ($1/n$)。

再举个例子，鸢尾花数据集的 150 个样本均分为 3 类——*setosa* (C_1)、*versicolour* (C_2)、*virginica* (C_3)。如图 5 所示，从 150 个样本中取出任一样本，样本标签为 C_1 、 C_2 、 C_3 对应的概率相同，都是：

$$\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \frac{50}{150} = \frac{1}{3} \quad (5)$$

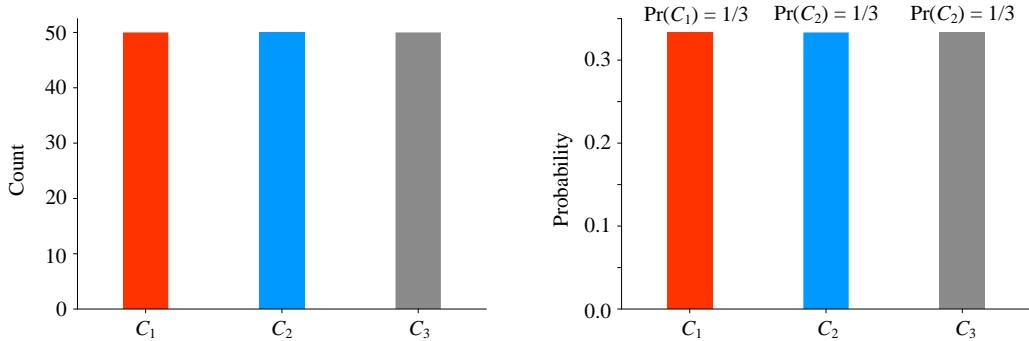


图 5. 鸢尾花 150 个样本数据均分为三类

抛一枚硬币

抛一枚硬币，1 代表正面，0 代表反面。抛一枚硬币可能结果的样本空间为：

$$\Omega = \{0, 1\} \quad (6)$$

假设硬币质地均匀，获得正面和反面的概率相同，均为 $1/2$ ，即：

$$\Pr(0) = \Pr(1) = \frac{1}{2} \quad (7)$$

把 $\{0, 1\}$ 标记在数轴上，用火柴梗图可视化上述概率值，我们便得到图 6。

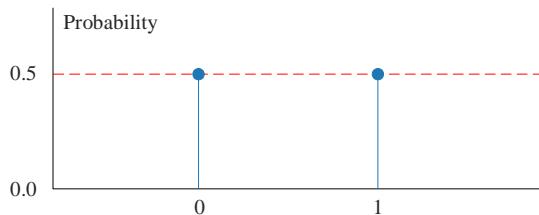


图 6. 抛一枚硬币结果和对应的理论概率值

图 7 所示为反复抛一枚硬币，正面 (1)、反面 (0) 平均值随试验次数变化。可以发现平均结果不断靠近 $1/2$ ，也就是说正反面出现的概率几乎相同。

从另外一个角度，(7) 给出的是用古典概率模型（等可能事件和枚举法）得出的**理论概率**（theoretical probability）。也称为公式概率或数学概率，是一种基于理论推导的概率计算方法。它一般基于假设所有可能的结果是等可能的，并使用数学公式计算概率。

而图 7 是采用试验得到的统计结果，印证了概率模型结果。根据大量的、重复的统计试验结果计算随机事件中各种可能发生结果的概率，称为**试验概率** (experimental probability)。试验概率是一种基于实际试验的概率计算方法。它通过多次重复试验来统计某个事件发生的频率，然后将频率作为概率的估计值。

理论概率可以作为试验概率的基础，即在假设所有可能的结果是等可能的情况下，理论概率可以预测事件发生的概率，而试验概率则可以验证这一预测是否准确。



本书第 15 章介绍如何完成蒙特卡罗模拟 (Monte Carlo simulation)。

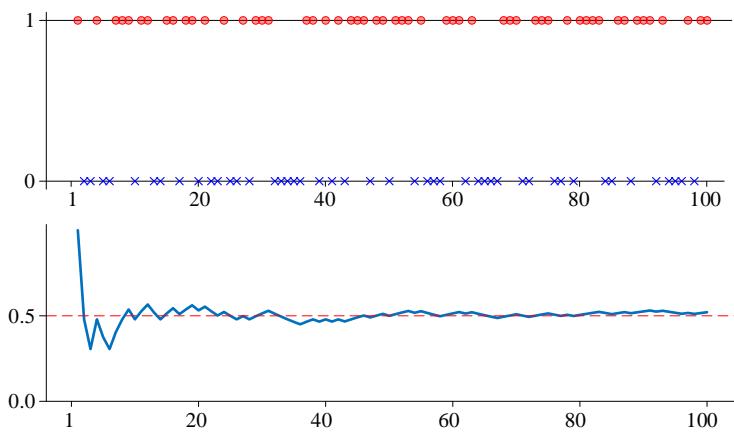


图 7. 抛硬币 100 次试验结果变化

掷色子

如图 8 所示，掷一枚色子试验可能结果的样本空间为：

$$\Omega = \{1, 2, 3, 4, 5, 6\} \quad (8)$$

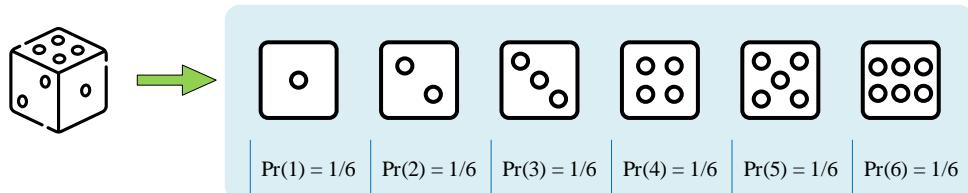


图 8. 投色子试验

试验中，假设获得每一种点数的可能性相同。掷一枚色子共 6 种结果，每种结果对应的概率为：

$$\Pr(1) = \Pr(2) = \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = \frac{1}{6} \quad (9)$$

同样用火柴梗图把上述结果画出来，得到图 9。这也是抛一枚色子得到不同点数对应概率的理论值。

然而实际情况可能并非如此。想象一种特殊情况，某一枚特殊的色子，它的质地不均匀，可能产生点数 6 的概率略高于其他点数。这种情况下，要想估算不同结果的概率值，一般只能通过试验。

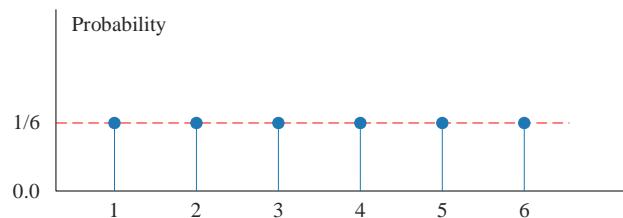


图 9. 抛一枚色子结果和对应的理论概率值

抛两枚硬币

下面看两个稍复杂的例子——每次抛两枚硬币。

比如，如果第一枚硬币正面、第二枚硬币反面，结果记做 $(1, 0)$ 。这样，样本空间由以下 4 个点构成：

$$\Omega = \{(0,0), (0,1), (1,0), (1,1)\} \quad (10)$$

图 10 (a) 所示为用二维坐标系展示试验结果。图中横轴代表第一枚硬币点数，纵轴为第二枚硬币对应点数。

假设，两枚硬币质地均匀，抛一枚硬币获得正、反面的概率均为 $1/2$ 。而抛两枚硬币对应结果的概率如图 10 (b) 所示。

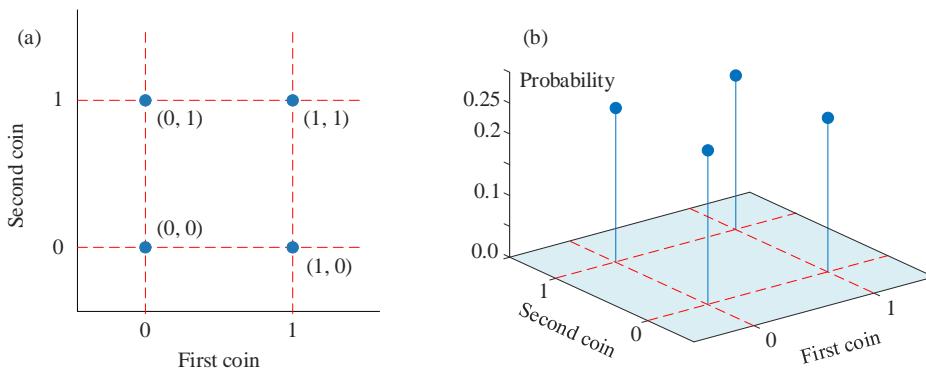


图 10. 抛两枚硬币结果和对应的理论概率值

抛两枚色子

同理，每次抛 2 枚色子，样本空间 Ω 的等可能试验结果数量为 6×6 ：

$$\Omega = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\} \quad (11)$$

图 11 (a) 所示为上述试验的样本空间。图 11 (b) 中，假设色子质地均匀，每个试验结果对应的概率均为 $1/36$ 。

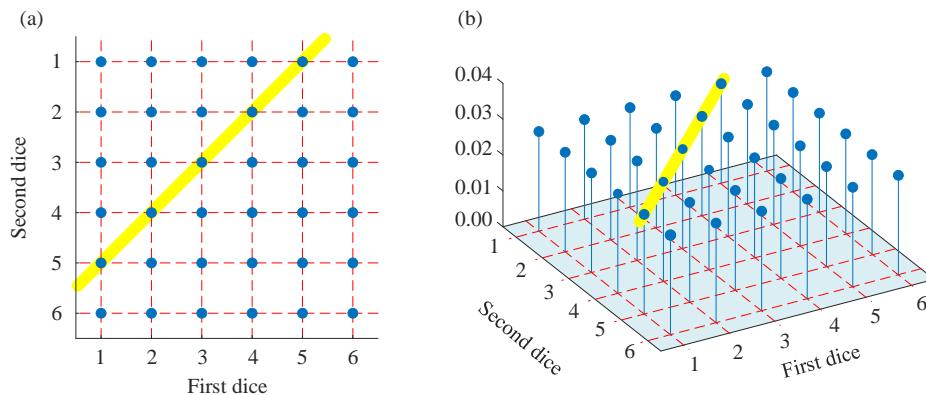


图 11. 抛两枚色子结果和对应的理论概率值

抛两枚色子：点数之和为 6

下面，我们看一种特殊情况。如图 12 所示，如果我们关心两个色子点数之和为 6 的话，发现一共有五种结果满足条件。这五种结果为 $1+5$ 、 $2+4$ 、 $3+3$ 、 $4+2$ 、 $5+1$ 。该事件对应概率为：

$$\Pr(\text{sum} = 6) = \frac{5}{6 \times 6} \approx 0.1389 \quad (12)$$

图 11 (a) 中黄色背景所示样本便代表抛两枚色子点数之和为 6 的事件。

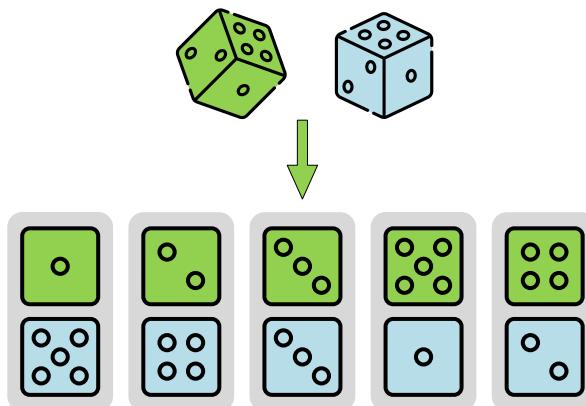


图 12. 投两个色子，点数之和为 6

编写代码进行 10,000,000 次试验，累计“点数之和为 6”事件发生次数，并且计算该事件当前概率。图 13 所示“点数之和为 6”事件概率随抛掷次数变化曲线。

比较 (12) 和图 13，通过古典概率模型得到的理论结论和试验结果相互印证。



图 13 横轴为对数刻度。《数学要素》第 12 章介绍过对数刻度，大家可以回顾。

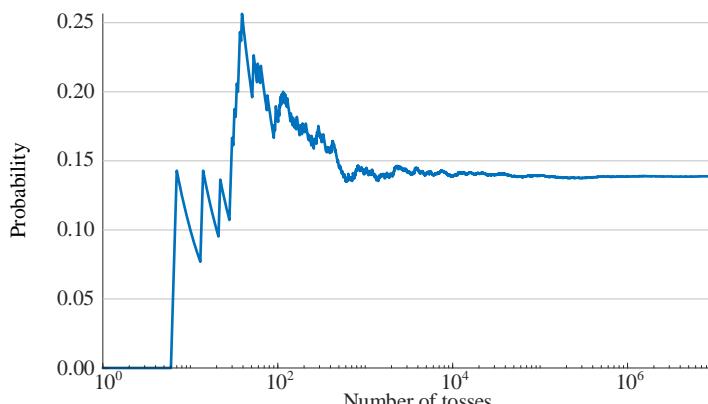
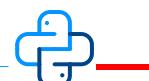


图 13. “色子点数之和为 6”事件概率随抛掷次数变化



代码 Bk5_Ch03_01.py 模拟抛色子试验并绘制图 13。请大家把这个代码改写成一个 Streamlit App，并用抛掷次数作为输入。

抛两枚色子：点数之和的样本空间

接着上一个例子，如果我们要对抛两枚色子“点数之和”感兴趣，首先要知道这个事件的样本空间。如图 14 所示，彩色等高线对应两枚色子点数之和。由此，得到两个色子点数之和的样本空间为 {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}。

而等高线上灰色点 ● 的横纵坐标代表满足条件的色子点数。计算某一条等高线上点 ● 的数量，再除 $36 (= 6 \times 6)$ 便得到不同“点数之和”对应的概率值。

图 14 (b) 所示样本空间所有结果概率值的火柴梗图。观察图 14 (b)，容易发现结果非等概率；但是，这些概率值也是通过等概率模型推导得到。

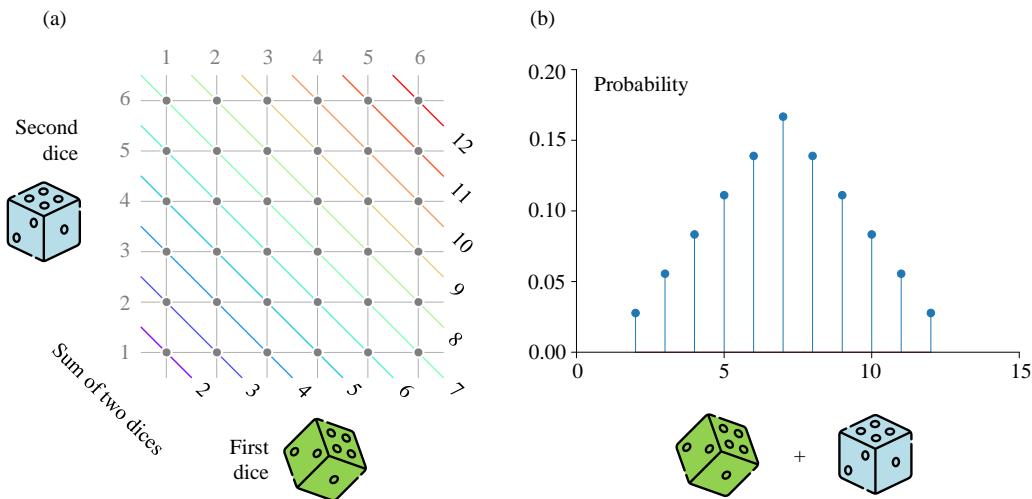


图 14. 两个色子点数之和

更多“花样”

接着上面抛两枚色子算点数之和的试验，我们玩出更多“花样”！

如表 1 所示，抛两枚色子，我们可以只考虑第一只色子的点数、第一只色子点数平方值，也可以计算两个色子的点数平均值、乘积、商、差、差的平方等等。

这些不同的花式玩法至少告诉我们以下几层信息：

- ▶ 抛两枚色子，第一枚色子和第二枚色子的结果可以独立讨论；换个视角来看，一次试验中，第一、二枚色子点数结果相互不影响；
- ▶ 第一枚和第二枚色子的点数结果还可以继续运算；
- ▶ 用文字描述这些结果太麻烦了，我们需要将它们代数化！比如，定义第一个色子结果为 X_1 ，第二个色子点数为 X_2 ，两个点数数学运算结果为 Y 。这便是下一章要探讨的**随机变量** (random variable)。
- ▶ 显然表 1 中每种花式玩法有各自的样本空间 Ω 。样本空间的样本并非都是等概率。但是，样本空间中所有样本的概率之和都是 1。

→ 表 1 所示为基于抛两枚色子试验结果的更多花式玩法。请大家试着找到每种运算的样本空间，并计算每个样本对应的概率值。我们将在下一章揭晓答案。

表 1. 基于抛两枚色子试验结果的更多花式玩法

随机变量	描述	例子											
X_1	第一个色子点数	1	2	3	4	5	6	1	2	3	4	5	6
X_2	第二个色子点数	1	1	1	1	1	1	2	2	2	2	2	2
$Y = X_1$	只考虑第一个色子点数	1	2	3	4	5	6	1	2	3	4	5	6
$Y = X_1^2$	第一个色子点数平方	1	4	9	16	25	36	1	4	9	16	25	36
$Y = X_1 + X_2$	点数之和	2	3	4	5	6	7	3	4	5	6	7	8
$Y = \frac{X_1 + X_2}{2}$	点数平均值	1	1.5	2	2.5	3	3.5	1.5	2	2.5	3	3.5	4
$Y = \frac{X_1 + X_2 - 7}{2}$	中心化点数之和，再求平均	-2.5	-2	-1.5	-1	-0.5	0	-2	-1.5	-1	-0.5	0	0.5
$Y = X_1 X_2$	点数之积	1	2	3	4	5	6	2	4	6	8	10	12
$Y = \frac{X_1}{X_2}$	点数之商	1	2	3	4	5	6	0.5	1	1.5	2	2.5	3
$Y = X_1 - X_2$	点数之差	0	1	2	3	4	5	-1	0	1	2	3	4
$Y = X_1 - X_2 $	点数之差的绝对值	0	1	2	3	4	5	1	0	1	2	3	4
$Y = (X_1 - 3.5)^2 + (X_2 - 3.5)^2$	中心化点数平方和	12.5	8.5	6.5	6.5	8.5	12.5	8.5	4.5	2.5	2.5	4.5	8.5

抛三枚色子

为了大家习惯“多元”思维，我们再进一步将一次抛掷色子的数量提高至三枚。

第一枚点数定义为 X_1 ，第二枚 X_2 ，第三枚 X_3 。

图 15 (a) 所示为抛三枚色子点数的样本空间，这显然是个三维空间。比如，坐标点 $(3, 3, 3)$ 代表三枚色子的点数都是 3。

图 15 (a) 这个样本空间有 $216 (= 6 \times 6 \times 6)$ 个样本。假设这三个色子质量均匀，获得每个点数为等概率，则图 15 (a) 中每个样本对应的概率为 $1/216$ 。

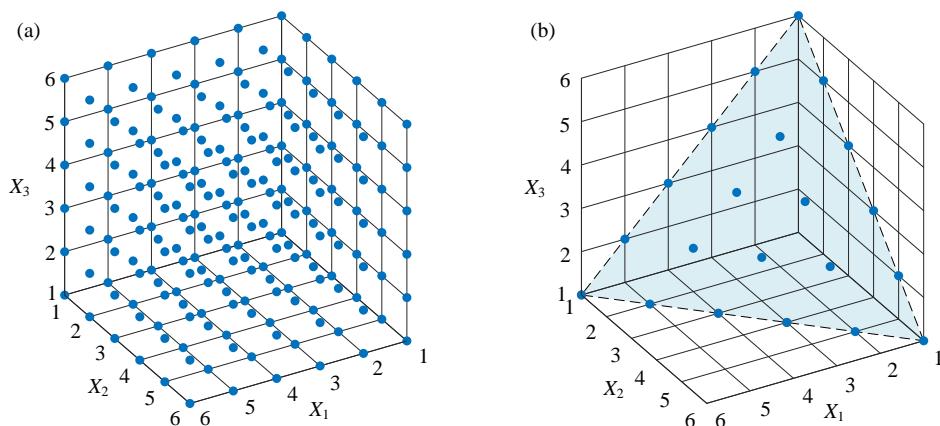


图 15. 抛三枚色子点数的样本空间

定义事件 A 为三枚色子的点数之和为 8，即 $X_1 + X_2 + X_3 = 8$ 。事件 A 对应的样本集合如图 15 (b) 所示，一共有 21 个样本点，容易发现这些样本在同一个斜面上。相对图 15 (a) 这个样本空间，事件 A 的概率为 $21/216$ 。

大家可能已经发现，实际上，我们可以用水平面来可视化事件 A 的样本集合。如图 16 所示，将散点投影在平面上得到图 16 (b)。能够完成这种投影是因为 $X_1 + X_2 + X_3 = 8$ 这个等式关系。

通过这个例子，大家已经发现多元统计中，几何思维的重要性。



这种投影思路将会用到本书后续要介绍的多项分布（第 5 章）、Dirichlet 分布（第 7 章）。

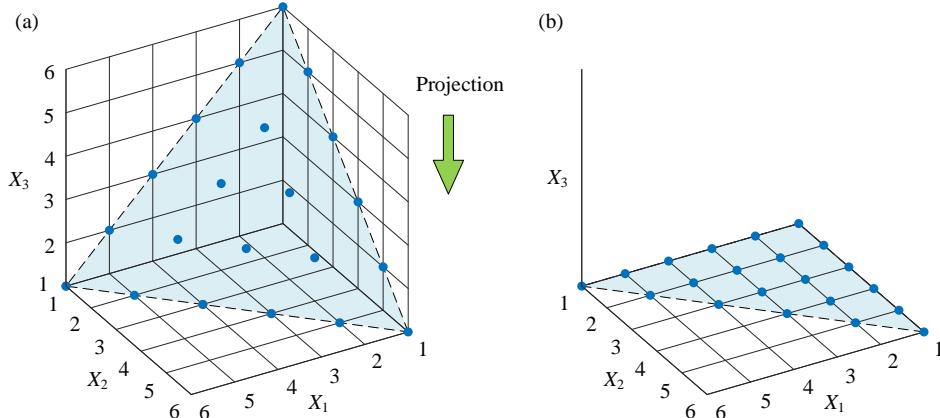


图 16. 将事件 A 的样本点投影到平面上

3.3 回顾：杨辉三角和概率

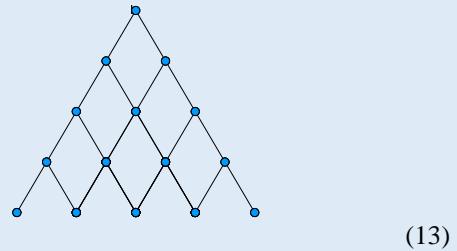
杨辉三角



《数学要素》第 20 章介绍过杨辉三角和古典概率模型的联系，本节稍作回顾。

杨辉三角又叫**帕斯卡三角** (Pascal's triangle)，是二项式系数的一种写法。 $(a + b)^n$ 展开后，按单项 a 的次数从高到低排列得到：

$$\begin{aligned}
 (a+b)^0 &= 1 \\
 (a+b)^1 &= a+b \\
 (a+b)^2 &= a^2 + 2ab + b^2 \\
 (a+b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3 \\
 (a+b)^4 &= a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4
 \end{aligned}$$



(13)

其中， a 和 b 均不为 0。

抛硬币

把二项式展开用在理解抛硬币的试验。 $(a+b)^n$ 中 n 代表一次抛掷中硬币数量， a 可以理解为“硬币正面朝上”对应概率， b 为“硬币反面朝上”对应概率。如果硬币质地均匀， $a = b = 1/2$ 。

举个例子，如果硬币质地均匀，每次抛 10 (n) 枚硬币，正好出现 6 次正面对应概率为：

$$\Pr(\text{heads} = 6) = C_{10}^6 \frac{1}{2^{10}} = \frac{210}{1024} = \frac{210}{1024} \approx 0.20508 \quad (14)$$

每次抛 10 枚硬币，至少出现 6 次正面的概率为：

$$\Pr(\text{heads} \geq 6) = \frac{C_{10}^6 + C_{10}^7 + C_{10}^8 + C_{10}^9 + C_{10}^{10}}{2^{10}} = \frac{210 + 120 + 45 + 10 + 1}{1024} = \frac{386}{1024} \approx 0.37695 \quad (15)$$

编写代码，一共抛 10000 次，每次抛 10 枚硬币。分别累计“正好出现 6 次正面”、“至少出现 6 次正面”两个事件的次数，并且计算两个事件当前概率。图 17 所示两事件概率随抛掷次数变化曲线。这也是试验概率、理论概率相互印证。

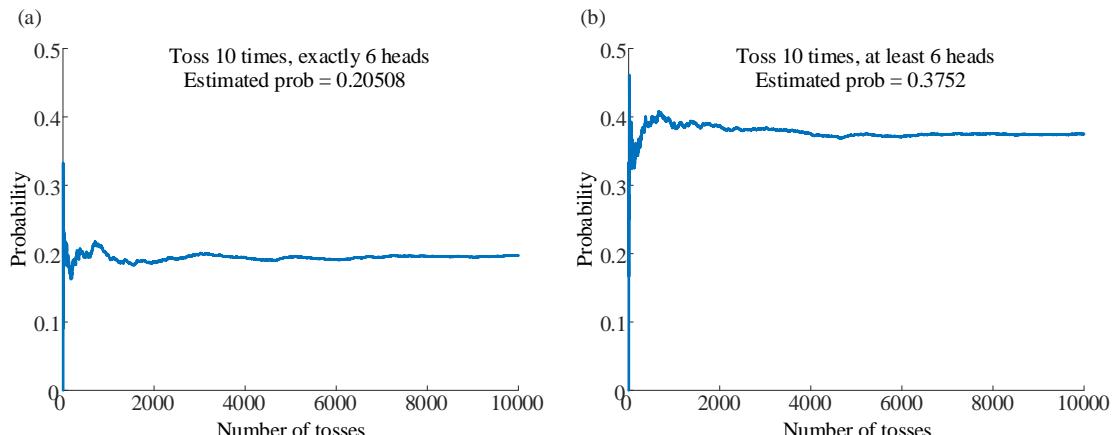


图 17. 试验概率随抛掷次数变化：a) 正好出现 6 次正面；b) 至少出现 6 次正面



Bk5_Ch03_02.py 完成上述两个试验并绘制图 17。

回忆二叉树

《数学要素》第 20 章还介绍过杨辉三角和二叉树的联系，如图 18 所示。

站在二叉树中间节点处，向上走、还是向下走对应的概率便分别对应“硬币正面朝上”、“硬币反面朝上”概率。

假设，向上走、向下走的概率均为 $1/2$ 。图 18 右侧的直方图展示了两组数，分别是达到终点不同节点的路径数量、概率值。请大家回忆如何用组合数计算这些概率值。

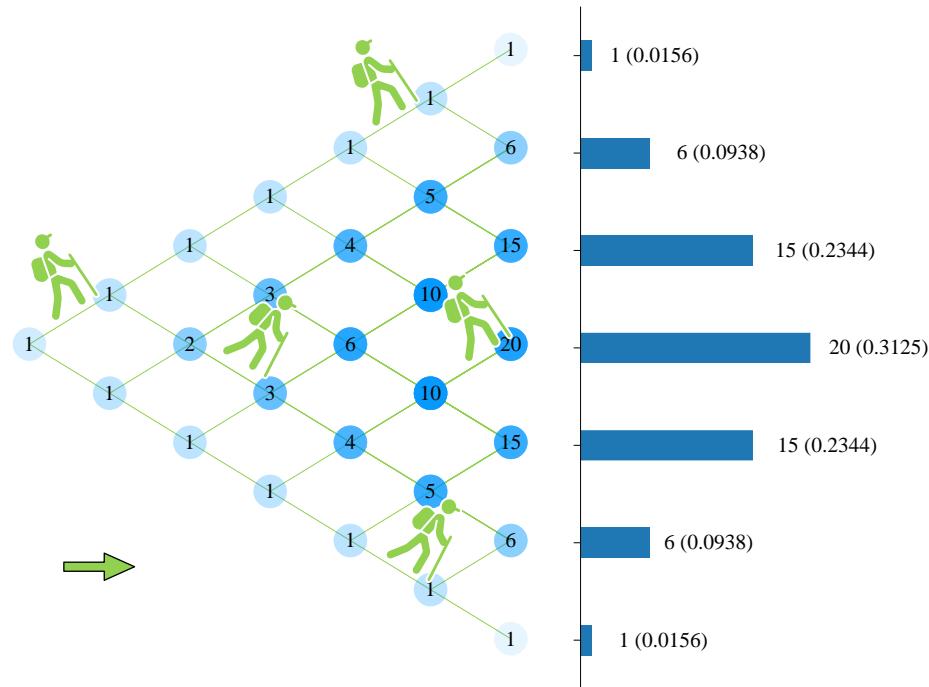


图 18. 杨辉三角逆时针旋转 90 度得到一个二叉树，图片基于《数学要素》第 20 章

3.4 事件之间的关系：集合运算

积事件

事件 A 与事件 B 为样本空间 Ω 中的两个事件， $A \cap B$ 代表 A 和 B 的**积事件** (the intersection of events A and B)，指的是某次试验时，事件 A 和事件 B 同时发生。

$\Pr(A \cap B)$ 代表 A 和 B **积事件概率** (probability of the intersection of events A and B 或 joint probability of A and B)。 $\Pr(A \cap B)$ 也叫做 A 和 B **联合概率** (joint probability)。 $\Pr(A \cap B)$ 也常记做 $\Pr(A, B)$ ：

$$\Pr_{\text{Joint}}(A \cap B) = \Pr_{\text{Joint}}(A, B)$$

(16)

互斥

如果事件 A 与事件 B 为两者交集为空 $A \cap B = \emptyset$ ，则称**事件 A 和事件 B 互斥** (events A and B are disjoint)，或称 **A 和 B 互不相容** (two events are mutually exclusive)。

白话说，事件 A 与事件 B 不可能同时发生，也就是说 $\Pr(A \cap B)$ 为 0：

$$A \cap B = \emptyset \Rightarrow \Pr_{\text{Joint}}(A \cap B) = \Pr_{\text{Joint}}(A, B) = 0$$

(17)

和事件

事件 $A \cup B$ 为 A 和 B 的**和事件** (union of events A and B)。具体来说，当事件 A 和事件 B 至少有一个发生时，事件 $A \cup B$ 发生。 $\Pr(A \cup B)$ 代表事件 A 和 B **和事件概率** (probability of the union of events A and B 或 probability of A or B)。

$\Pr(A \cup B)$ 和 $\Pr(A \cap B)$ 之间关系为：

$$\underbrace{\Pr(A \cup B)}_{\text{Union}} = \Pr(A) + \Pr(B) - \underbrace{\Pr(A \cap B)}_{\text{Joint}}$$

(18)

如果事件 **A 和 B 互斥** (events A and B are mutually exclusive)，即 $A \cap B = \emptyset$ 。对于这种特殊情况， $\Pr(A \cup B)$ 为：

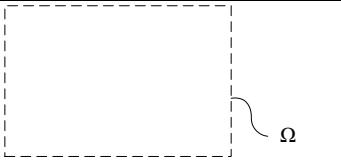
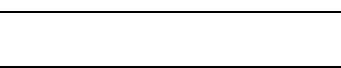
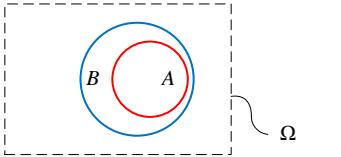
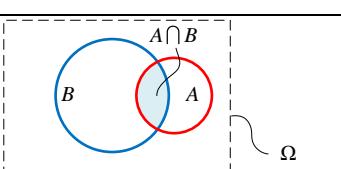
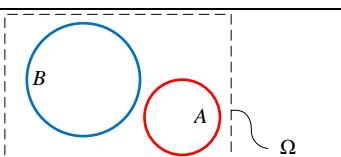
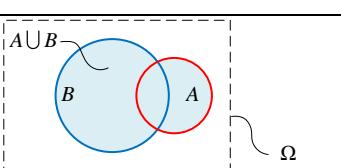
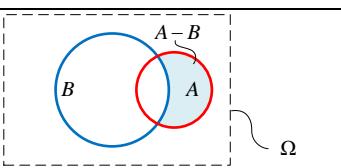
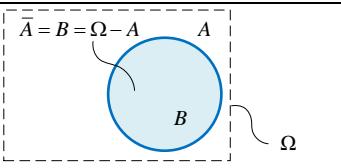
$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

(19)

表 2 总结常见集合运算维恩图。

表 2. 常见集合运算和维恩图

符号	解释	维恩图

Ω	必然事件，即整个样本空间 (sample space)	
\emptyset	不可能事件，即空集 (empty set)	
$A \subset B$	事件 B 包含事件 A (event A is a subset of event B) 即，事件 A 发生，事件 B 必然发生	
$A \cap B$	事件 A 和事件 B 的积事件 (the intersection of events A and B) 即，某次试验时，当事件 A 和事件 B 同时发生时，事件 $A \cap B$ 发生	
$A \cap B = \emptyset$	事件 A 和事件 B 互斥 (events A and B are disjoint)，两个事件互不相容 (two events are mutually exclusive) 即，事件 A 和事件 B 不能同时发生	
$A \cup B$	事件 A 和事件 B 的和事件 (the union of events A and B) 即，当事件 A 和事件 B 至少有一个发生时，事件 $A \cup B$ 发生	
$A - B$	事件 A 与事件 B 的差事件 (the difference between two events A and B) 即，事件 A 发生、事件 B 不发生， $A - B$ 发生	
$A \cup B = \Omega$ 且 $A \cap B = \emptyset$ 也可以记做 $\bar{A} = B = \Omega - A$ (complement of event A)	事件 A 与事件 B 互为逆事件 (complementary events)，对立事件 (collectively exhaustive) 即，对于任意一次试验，事件 A 和事件 B 有且仅有一个发生	

3.5 条件概率：给定部分信息做推断

条件概率 (conditional probability) 是在给定部分信息基础上对试验结果的一种推断。条件概率是机器学习、数学科学中至关重要概念，本书大多数内容都是围绕条件概率展开，请大家格外留意。

三个例子

下面给出三个例子说明哪里会用到“条件概率”。

在抛两个色子试验中，事件 A 为其中一个色子点数为 5，事件 B 为点数之和为 6。给定事件 B 发生条件下，事件 A 发生的概率多少？

给定花萼长度为 5 厘米，花萼宽度为 2 厘米。根据 150 个鸢尾花样本数据，鸢尾花样本最可能是哪一类 (setosa、versicolor、virginica)？对应的概率大概是多少？

根据 150 个鸢尾花样本数据，如果某一朵鸢尾花的花萼长度为 5 厘米，它的花萼宽度最可能多宽？

条件概率

A 和 B 为样本空间 Ω 中的两个事件，其中 $\Pr(B) > 0$ 。那么，**事件 B 发生的条件下事件 A 发生的条件概率** (conditional probability of event A occurring given B occurs 或 probability of A given B) 可以通过下式计算得到：

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

(20)

其中， $\Pr(A \cap B)$ 为 A 和 B 事件的联合概率， $\Pr(B)$ 也叫 B 事件边缘概率。

⚠ 注意，我们也可以这么理解 $\Pr(A|B)$ ， B 实际上是“新的样本空间”—— Ω_B ！ $\Pr(A|B)$ 是在 Ω_B 中计算得到的概率值。

$\Pr(B)$ 、 $\Pr(A \cap B)$ 都是在 Ω 中计算得到的概率值。

Ω_B 是的子集 Ω ，两者的联系正是 $\Pr(B)$ ，即 B 在 Ω 中对应的概率。 $\Pr(B)$ 也可以写成“条件概率”的形式 $\Pr(B|\Omega)$ 。

类似地，事件 A 发生的条件下事件 B 发生的条件概率为：

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}$$

(21)

其中， $\Pr(A)$ 为 A 事件边缘概率， $\Pr(A) > 0$ 。

类似地， $\Pr(B|A)$ 也可以理解为 B 在“新的样本空间” Ω_A 中的概率。

联合概率

利用(20), 联合概率 $\Pr(A \cap B)$ 可以整理为：

$$\underbrace{\Pr(A \cap B)}_{\text{Joint}} = \underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(A|B)}_{\text{Conditional}} \cdot \underbrace{\Pr(B)}_{\text{Marginal}}$$

(22)

上式相当于“套娃”。首先在 Ω_B 中考虑 A (实际上是 $A \cap B$)，然后把 $A \cap B$ 再放回 Ω 中。也就是说，把 $\Pr(A|B)$ 写成 $\Pr(A \cap B|B)$ 也没问题。因为， A 只有 $A \cap B$ 这部分在 $B(\Omega_B)$ 中。

同样， $\Pr(A \cap B)$ 也可以写成：

$$\underbrace{\Pr(A \cap B)}_{\text{Joint}} = \underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(B|A)}_{\text{Conditional}} \underbrace{\Pr(A)}_{\text{Marginal}}$$

(23)

举个例子

掷一颗色子，一共有 6 种等概率结果 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 。

事件 B 为“点数为奇数”，事件 C 为“点数小于 4”。事件 B 的概率 $\Pr(B) = 1/2$ ，事件 C 的概率 $\Pr(C) = 1/2$ 。

如图 19 所示， $B \cap C$ 事件发生的概率 $\Pr(B \cap C) = \Pr(B, C) = 1/3$ 。

在事件 B (点数为奇数) 条件下，事件 C (点数小于 4) 发生的条件概率为：

$$\Pr(C|B) = \frac{\Pr(B \cap C)}{\Pr(B)} = \frac{\Pr(B, C)}{\Pr(B)} = \frac{1/3}{1/2} = \frac{2}{3}$$

(24)

图 19 也告诉我们一样的结果。请大家回顾本章最初给出孟德尔豌豆试验和道尔顿红绿色盲，手算其中的条件概率。

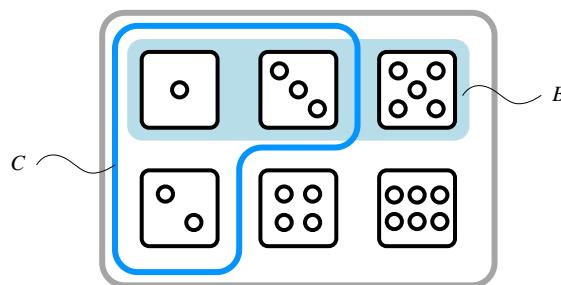


图 19. 事件 B 发生条件下事件 C 发生的条件概率

推广

(22) 可以继续推广， A_1, A_2, \dots, A_n 为 n 个事件，它们的联合概率可以展开写成一系列条件概率的乘积：

$$\begin{aligned} \Pr(A_1 \cap A_2 \cap \dots \cap A_n) &= \Pr(A_1, A_2, A_3, \dots, A_{n-1}, A_n) \\ &= \Pr(A_n | A_1, A_2, A_3, \dots, A_{n-1}) \Pr(A_{n-1} | A_1, A_2, A_3, \dots, A_{n-2}) \dots \Pr(A_2 | A_1) \Pr(A_1) \end{aligned} \quad (25)$$

这也叫做条件概率的**链式法则** (chain rule)。

比如， $n = 4$ 时，上式可以写成：

$$\begin{aligned} \underbrace{\Pr(A_1, A_2, A_3, A_4)}_{\text{Joint}} &= \underbrace{\Pr(A_4 | A_1, A_2, A_3)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_1, A_2, A_3)}_{\text{Joint}} \\ &= \underbrace{\Pr(A_4 | A_1, A_2, A_3)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_3 | A_1, A_2)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_1, A_2)}_{\text{Joint}} \\ &= \underbrace{\Pr(A_4 | A_1, A_2, A_3)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_3 | A_1, A_2)}_{\text{Conditional}} \cdot \underbrace{\Pr(A_2 | A_1)}_{\text{Conditional}} \Pr(A_1) \end{aligned} \quad (26)$$


大家可以把上式想成多层套娃。上式配图假设事件相互之间完全包含，这样方便理解。实际上，事件求积的过程已经将“多余”的部分切掉：



$$(A_1 \cap A_2 \cap A_3 \cap A_4) \subset (A_1 \cap A_2 \cap A_3) \subset (A_1 \cap A_2) \subset A_1 \quad (27)$$

3.6 贝叶斯定理：条件概率、边缘概率、联合概率关系

贝叶斯定理 (Bayes' theorem) 是由**托马斯·贝叶斯** (Thomas Bayes) 提出。毫不夸张地说，贝叶斯定理撑起机器学习、深度学习算法的半边天。

贝叶斯定理的基本思想是根据**先验概率** (prior) 和新的**证据** (evidence) 来计算**后验概率** (posterior)。在实际应用中，我们通常根据一些已知的先验知识，来计算事件的先验概率。然后，当我们获取新的证据时，就可以利用贝叶斯定理来计算事件的后验概率，从而更新我们的信念或概率。

→ 本书后续将见缝插针地讲解贝叶斯定理和应用，特别是在贝叶斯分类 (第 18、19 两章)、贝叶斯推断 (第 20、21、22 三章) 两个话题中。

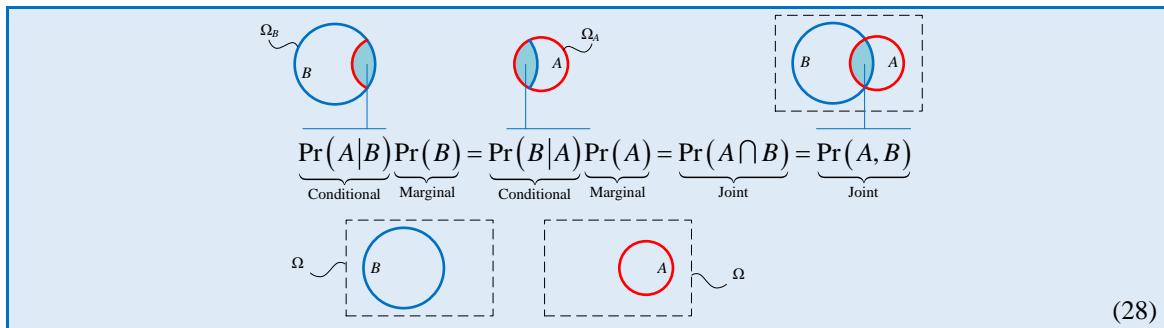


托马斯·贝叶斯 (Thomas Bayes) | 英国数学家 | 1702 ~ 1761

贝叶斯统计的开山鼻祖，以贝叶斯定理闻名于世。

关键词：● 贝叶斯定理 ● 贝叶斯派 ● 贝叶斯推断 ● 朴素贝叶斯分类 ● 贝叶斯回归

贝叶斯定理描述的是两个条件概率的关系：



其中：

- ◀ $\Pr(A|B)$ 是指在 B 发生条件下 A 发生的**条件概率** (conditional probability); 也就是说, $\Pr(A|B)$ 的样本空间为 Ω_B ;
- ◀ $\Pr(B|A)$ 是指在 A 发生条件下 B 发生的条件概率; 也就是说, $\Pr(B|A)$ 的样本空间为 Ω_A ;
- ◀ $\Pr(A)$ 是 A 的**边缘概率** (marginal probability), 不考虑事件 B 的因素, 样本空间为 Ω ;
- ◀ $\Pr(B)$ 是 B 的边缘概率, 不考虑事件 A 的因素, 样本空间为 Ω ;
- ◀ $\Pr(A \cap B)$ 是事件 A 和 B 的联合概率, 样本空间为 Ω 。

图 20 给出理解贝叶斯原理的图解法。

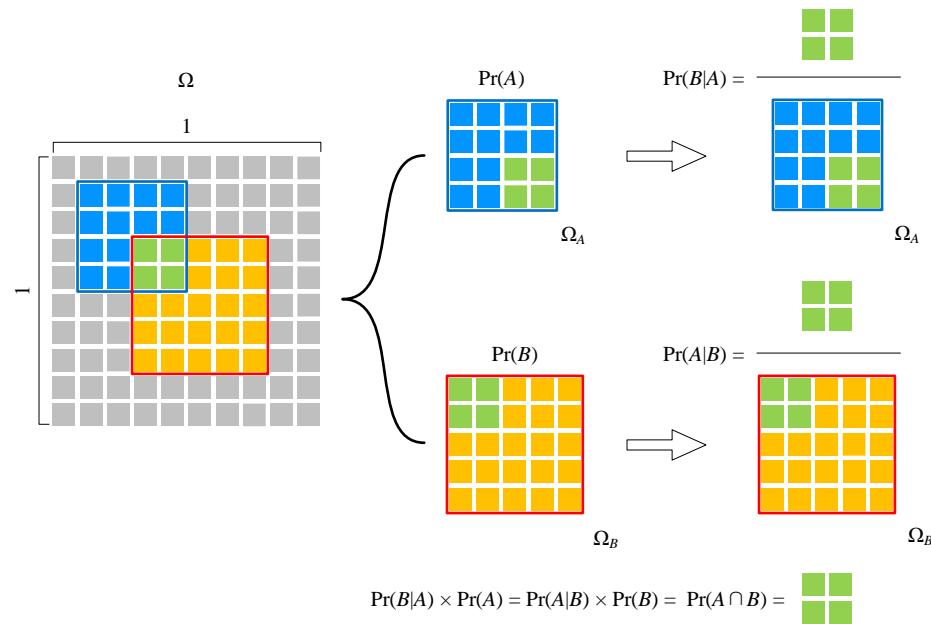


图 20. 贝叶斯原理图解

抛色子试验

现在，我们就用抛色子的试验来解释本节介绍的几个概率值。

根据本章前文内容，抛一枚色子可能得到 6 种结果，构成的样本空间为 $\Omega = \{1, 2, 3, 4, 5, 6\}$ 。假设每一种结果等概率，即 $\Pr(1) = \Pr(2) = \Pr(3) = \Pr(4) = \Pr(5) = \Pr(6) = 1/6$ 。

设“色子点数为偶数”事件为 A ，因此 $A = \{2, 4, 6\}$ ，对应概率为 $\Pr(A) = 3/6 = 0.5$ 。

A 事件的补集 B 对应事件“色子点数为奇数”， $B = \{1, 3, 5\}$ ，事件 B 的概率为 $\Pr(B) = 1 - \Pr(A) = 0.5$ 。

事件 A 和 B 交集 $A \cap B$ 为空集 \emptyset ，因此：

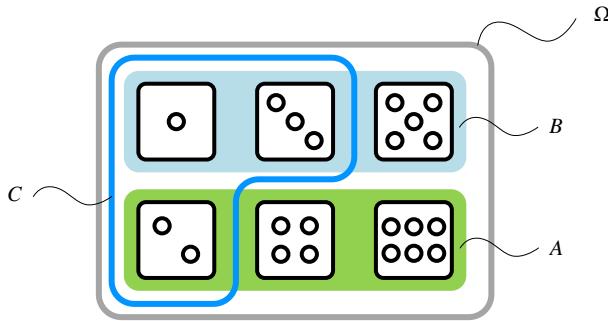
$$\Pr(A \cap B) = \Pr(A, B) = 0 \quad (29)$$

而 A 和 B 两者的并集 $A \cup B = \Omega$ ，因此对应的概率为 1：

$$\Pr(A \cup B) = 1 \quad (30)$$

C 事件被定为“色子点数小于 4”，因此 $C = \{1, 2, 3\}$ ，事件 C 的概率 $\Pr(C) = 0.5$ 。

图 22 展示的是 A 、 B 和 C 事件的关系。

图 21. A 、 B 、 C 事件定义

如图 22 (a) 所示，事件 A 和 C 的交集 $A \cap C = \{2\}$ ，因此 $A \cap C$ 的概率：

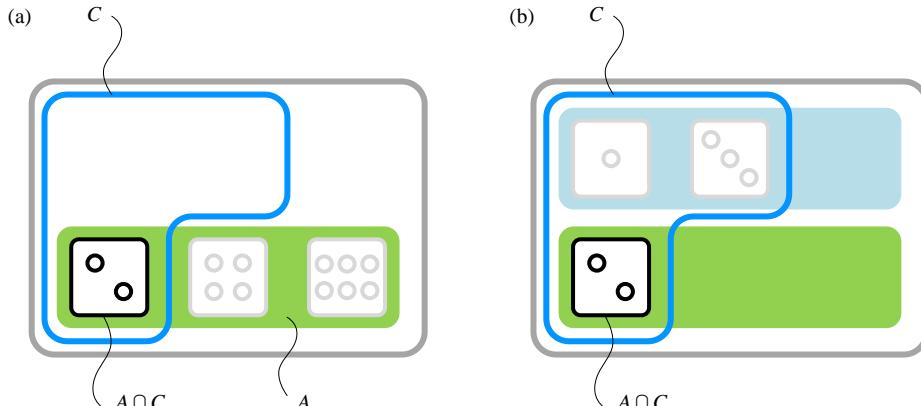
$$\Pr(A \cap C) = \Pr(A, C) = \frac{1}{6} \quad (31)$$

如图 22 (b) 所示，事件 B 和 C 的交集 $B \cap C = \{1, 3\}$ ，因此 $B \cap C$ 的概率：

$$\Pr(B \cap C) = \Pr(B, C) = \Pr(\{1\}) + \Pr(\{3\}) = \frac{1}{3} \quad (32)$$

A 和 C 的并集 $A \cup C = \{1, 2, 3, 4, 6\}$ ，对应的概率为：

$$\Pr(A \cup C) = \Pr(A) + \Pr(C) - \Pr(A, C) = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6} \quad (33)$$

图 22. 条件概率 $\Pr(C|A)$ 和条件概率 $\Pr(A|C)$

简单来说，条件概率 $\Pr(C|A)$ 代表在 A 事件发生的条件下， C 事件发生概率。用贝叶斯公式可以求解 $\Pr(C|A)$ ：

$$\Pr(C|A) = \frac{\Pr(A, C)}{\Pr(A)} = \frac{1/6}{1/2} = \frac{1}{3} \quad (34)$$

类似的，在 C 事件发生的条件下， A 事件发生的条件概率 $\Pr(A|C)$ 为：

$$\Pr(A|C) = \frac{\Pr(A, C)}{\Pr(C)} = \frac{1/6}{1/2} = \frac{1}{3} \quad (35)$$

请大家自行计算图 23 所示的 $\Pr(C|B)$ 和 $\Pr(B|C)$ 这两个条件概率。

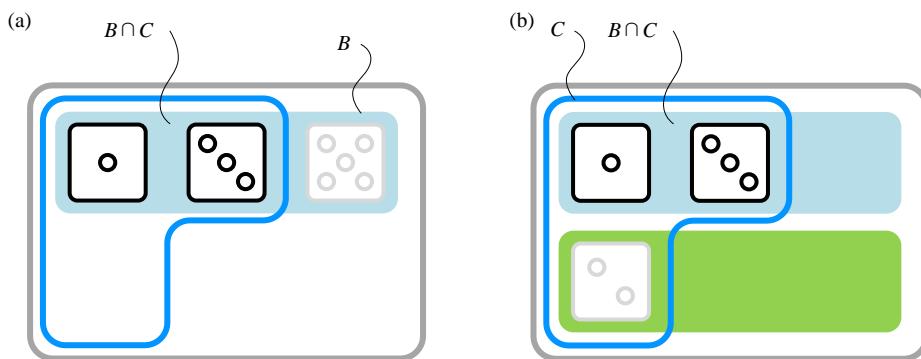


图 23. 条件概率 $\Pr(C|B)$ 和 $\Pr(B|C)$

贝叶斯定理是贝叶斯学派的核心工具。

频率学派 vs 贝叶斯学派

贝叶斯学派和频率学派是统计学中两种主要的哲学观点。它们之间的区别在于它们对概率的解释和使用方式不同。

频率学派将概率视为事件发生的频率或可能性，它强调基于大量数据和随机抽样的推断，通过检验假设来得出结论。频率学派侧重于经验数据和实证研究，常常使用假设检验和置信区间等方法来进行统计推断。

而贝叶斯学派则将概率视为一种个人信念的度量，它关注的是主观先验知识和经验的结合，以推断参数或未知量的后验分布。贝叶斯学派通常使用贝叶斯定理来计算后验分布，同时将不确定性视为一种核心特征，因此贝叶斯学派在处理小样本或缺乏数据的情况下表现更加优秀。

虽然贝叶斯学派和频率学派的基本理念和方法不同，但它们在某些情况下是相互补充的。例如，当样本数据较大时，频率学派的假设检验方法可以提供可靠的结果，而在缺乏数据或需要考虑主观经验和先验知识时，贝叶斯学派的方法则更为适用。此外，在一些实际应用中，两种方法可以相互结合，以得出更为准确的推断结论。

→ 本书后文将分别展开讲解频率学派(第 16、17 章)、贝叶斯学派(第 18~22 章)的应用场景。

3.7 全概率定理：穷举法

假设 A_1, A_2, \dots, A_n 互不相容，形成对样本空间 Ω 的分割 (partition)，也就是说每次试验事件 A_1, A_2, \dots, A_n 中有且仅有一个发生。

假定 $\Pr(A_i) > 0$ ，对于空间 Ω 中任意事件 B ，下式成立：

$$\begin{aligned}
 \Pr(B) &= \underbrace{\sum_{i=1}^n \Pr(A_i \cap B)}_{\text{Marginal}} = \Pr(A_1 \cap B) + \Pr(A_2 \cap B) + \dots + \Pr(A_n \cap B) \\
 &= \underbrace{\sum_{i=1}^n \Pr(A_i, B)}_{\text{Joint}} = \Pr(A_1, B) + \Pr(A_2, B) + \dots + \Pr(A_n, B)
 \end{aligned} \tag{36}$$

上式就叫做全概率定理 (law of total probability)。这本质上就是穷举法，也叫枚举法。

举个例子，图 24 给出的例子是三个互不相容事件 A_1, A_2, A_3 对 Ω 形成分割。通过全概率定理，即穷举法， $\Pr(B)$ 可以通过下式计算得到：

$$\Pr(B) = \underbrace{\Pr(A_1, B)}_{\text{Marginal}} + \underbrace{\Pr(A_2, B)}_{\text{Joint}} + \underbrace{\Pr(A_3, B)}_{\text{Joint}} \tag{37}$$

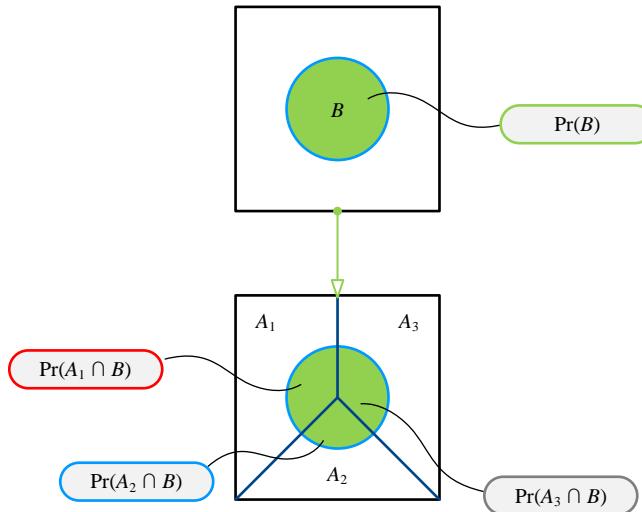


图 24. A_1, A_2, A_3 对空间 Ω 分割

引入贝叶斯定理

利用贝叶斯定理，以为 A_1, A_2, \dots, A_n 条件，展开 (36)：

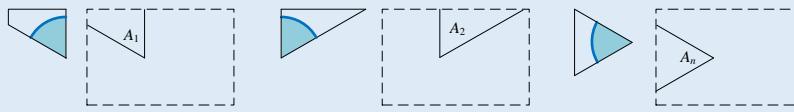
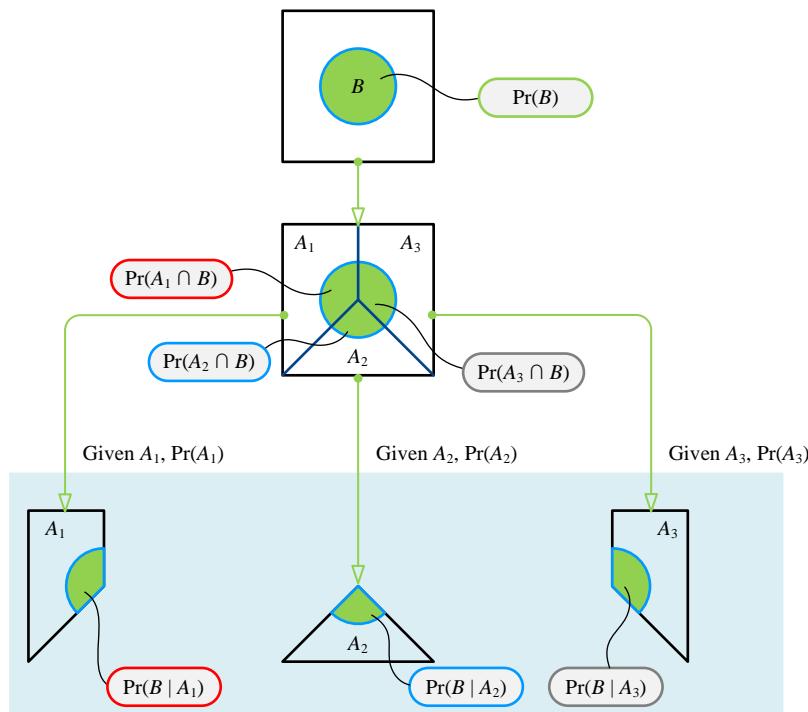
$$\begin{aligned}
 \Pr(B) &= \sum_{i=1}^n \underbrace{\Pr(A_i, B)}_{\text{Joint}} = \sum_{i=1}^n \underbrace{\Pr(B|A_i)}_{\text{Conditional}} \underbrace{\Pr(A_i)}_{\text{Marginal}} \\
 &= \Pr(B|A_1)\Pr(A_1) + \Pr(B|A_2)\Pr(A_2) + \cdots + \Pr(B|A_n)\Pr(A_n)
 \end{aligned}$$

(38)

图 25 所示为分别给定 A_1, A_2, A_3 条件下，事件 B 发生的情况。图 25. 分别给定 A_1, A_2, A_3 条件下，事件 B 发生的情况

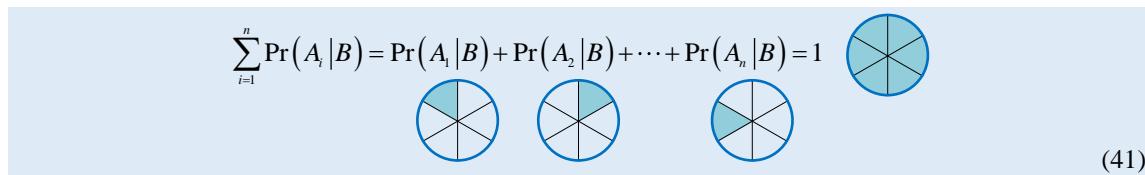
反过来，根据贝叶斯定理，在给定事件 B 发生条件下 ($\Pr(B) > 0$)，任意事件 A_i 发生的概率为：

$$\Pr(A_i|B) = \frac{\Pr(A_i, B)}{\Pr(B)} = \frac{\Pr(B|A_i) \cdot \Pr(A_i)}{\Pr(B)} \quad (39)$$

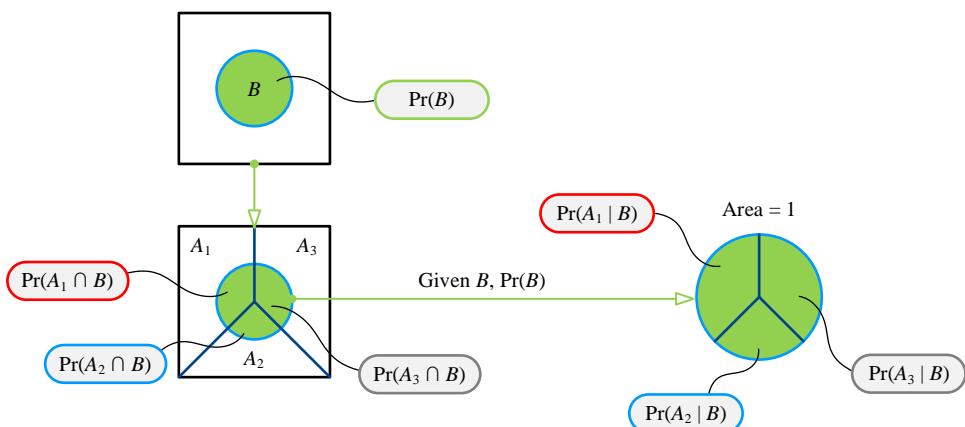
利用贝叶斯定理，以为 B 条件，进一步展开 (36)：

$$\begin{aligned}
 \Pr(B) &= \sum_{i=1}^n \underbrace{\Pr(A_i, B)}_{\text{Joint}} = \sum_{i=1}^n \underbrace{\Pr(A_i|B)}_{\text{Conditional}} \underbrace{\Pr(B)}_{\text{Marginal}} \\
 &= \Pr(A_1|B)\Pr(B) + \Pr(A_2|B)\Pr(B) + \cdots + \Pr(A_n|B)\Pr(B)
 \end{aligned}$$
(40)

(40) 等式左右消去 $\Pr(B)$ ($\Pr(B) > 0$)，得到：

图 26 所示为给定 B 条件下，事件 A_1 、 A_2 、 A_3 发生的情况。

看到这里，对贝叶斯定理和全概率定理还是一头雾水的读者不要怕，本书后续会利用不同实例反复讲解这两个定理。

图 26. 给定 B 条件下，事件 A_1 、 A_2 、 A_3 发生的情况

3.8 独立、互斥、条件独立

独立

上一节介绍的条件概率 $\Pr(A|B)$ 刻画了在事件 B 发生的条件下，事件 A 发生的可能性。

有一种特殊的情况，事件 B 发生与否，不会影响事件 A 发生的概率，也就是如下等式成立：

$$\underbrace{\Pr(A|B)}_{\text{Conditional}} = \underbrace{\Pr(A)}_{\text{Marginal}} \Leftrightarrow \underbrace{\Pr(B|A)}_{\text{Conditional}} = \underbrace{\Pr(B)}_{\text{Marginal}} \quad (42)$$

如果 (42) 给出的等式成立，则称**事件 A 和事件 B 独立** (events A and B are independent)。

如果 A 和 B 独立，联立 (28) 和 (42) 可以得到：

$$\Pr(A \cap B) = \underbrace{\Pr(A, B)}_{\text{Joint}} = \underbrace{\Pr(A)}_{\text{Marginal}} \cdot \underbrace{\Pr(B)}_{\text{Marginal}} \quad \begin{array}{c} \text{grid with 1 black cell} \\ \text{grid with 4 cells} \end{array} \quad (43)$$

如果一组事件 A_1 、 A_2 … A_n ，它们两两相互独立，则下式成立：

$$\Pr(A_1 \cap A_2 \cap \dots \cap A_n) = \Pr(A_1, A_2, \dots, A_n) = \Pr(A_1) \cdot \Pr(A_2) \cdots \Pr(A_n) = \prod_{i=1}^n \Pr(A_i) \quad (44)$$

抛三枚色子

接着本章前文“抛三枚色子”的例子。大家应该清楚，一次性抛三枚色子，这三枚色子点数互不影响，也就是“独立”。

如图 27 所示，第一枚色子的点数 (X_1) 取不同值 (1 ~ 6) 时，相当于把样本空间这个立方体切成 6 个“切片”。每个切片都有 36 个点，因此每个切片对应的概率均为：

$$\frac{6 \times 6}{6 \times 6 \times 6} = \frac{1}{6} \quad (45)$$

也就相当于把概率“1”，均分为 6 份。而 $1/6$ 对应第一枚色子的点数 (X_1) 取不同值的概率。

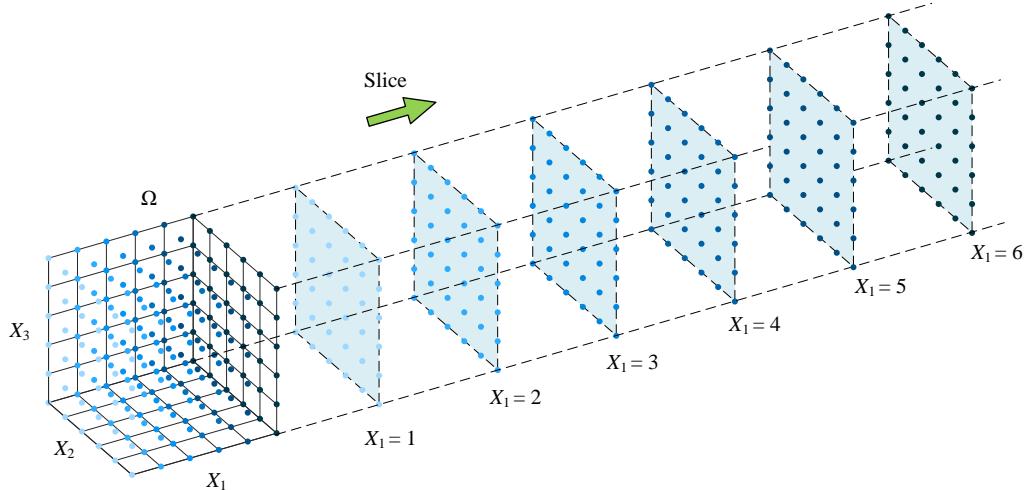


图 27. X_1 视角下的“抛三枚色子结果”

(3, 3, 3) 这个结果在整个样本空间中对应的概率为 $1/216$ 。如图 28 所示， $1/216$ 这个数值可以有四种不同的求法：

$$\frac{1}{216} = \frac{1}{6} \times \frac{1}{36} \quad (X_1=3) \quad (X_2, X_3)=(3,3) = \frac{1}{6} \times \frac{1}{36} \quad (X_2=3) \quad (X_1, X_3)=(3,3) = \frac{1}{6} \times \frac{1}{36} \quad (X_3=3) \quad (X_1, X_2)=(3,3) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \quad (X_1=3) \quad (X_2=3) \quad (X_3=3) \quad (46)$$

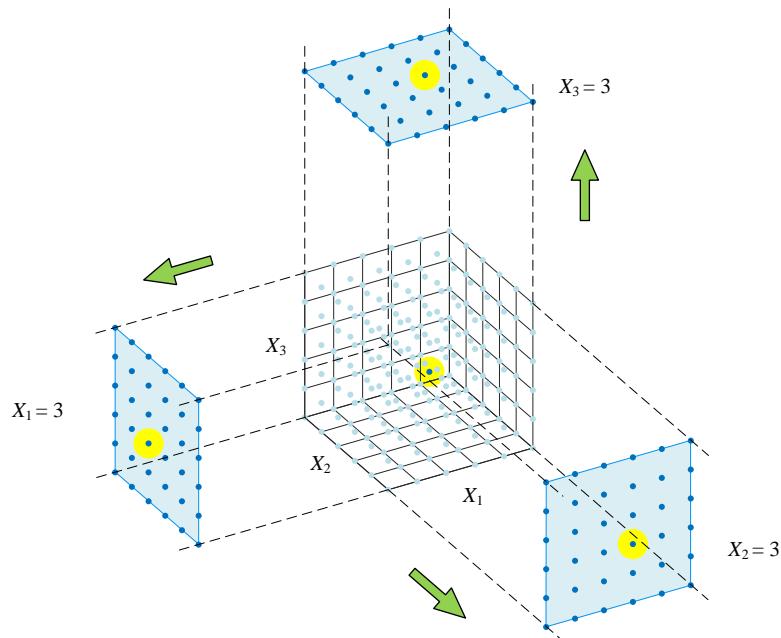


图 28. (3, 3, 3) 结果在样本空间和三个各方向切片上的位置

再换个角度，图 28 中立方体代表概率为 1，而 X_1 、 X_2 、 X_3 这三个随机变量独立，并将“1”均匀地切分成 216 份：

$$\left(\underbrace{\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}}_{X_1=1\sim 6} \right)^{-1} \times \left(\underbrace{\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}}_{X_2=1\sim 6} \right)^{-1} \times \left(\underbrace{\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}}_{X_3=1\sim 6} \right)^{-1} = 1 \quad (47)$$

上式体现的就是乘法分配律。从向量角度来看，上式相当于三个向量的张量积，撑起一个如图 28 所示的三维数组。

再次强调，之所以能用这种方式计算联合概率，就是因为三个随机变量“独立”。

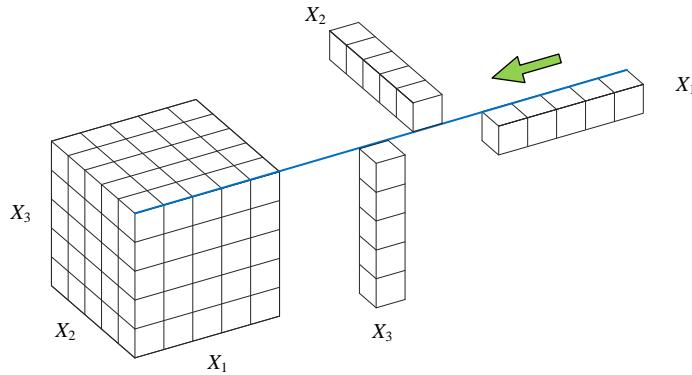


图 29. 三个向量的张量积

请大家格外注意，互斥不同于独立。[表 3](#) 对比一般情况、互斥、独立之间的主要特征。

[表 3. 比较一般情况、互斥、独立](#)

A 和 B	$\Pr(A \text{ and } B)$ $\Pr(A \cap B) = \Pr(A, B)$	$\Pr(A \text{ or } B)$ $\Pr(A \cup B)$	$\Pr(A B)$	$\Pr(B A)$
一般情况 $\Pr(A) > 0$ $\Pr(B) > 0$	$\Pr(A) \times \Pr(B A)$ $\Pr(B) \times \Pr(A B)$	$\Pr(A) + \Pr(B) - \Pr(A \cap B)$	$\Pr(A \cap B) / \Pr(B)$	$\Pr(A \cap B) / \Pr(A)$
互斥	0	$\Pr(A) + \Pr(B)$	0	0
独立	$\Pr(A) \times \Pr(B)$	$\Pr(A) + \Pr(B) - \Pr(A) \times \Pr(B)$	$\Pr(A)$	$\Pr(B)$

条件独立

在给定事件 C 发生条件下，如果如下等式成立，则称**事件 A 和事件 B 在 C 发生条件下条件独立** (events A and B are conditionally independent given an event C):

$$\Pr(A \cap B | C) = \Pr(A, B | C) = \Pr(A | C) \cdot \Pr(B | C) \quad (48)$$

⚠ 请大家格外注意， A 和 B 相互独立，无法推导得到 A 和 B 条件独立。而 A 和 B 条件独立，也无法推导得到 A 和 B 相互独立。本书后文还会深入讨论独立、条件独立。



古典概率有效地解决抛硬币、抛色子、口袋里摸球这些简单的概率问题，等概率模型、全概率定理、贝叶斯定理等重要的概率概念也随之产生。随着研究不断深入，概率统计工具的应用场景也开始变得更加多样。

基于集合论的古典概率模型渐渐地显得力不从心。引入随机变量、概率分布等概念，实际上就是将代数思想引入概率统计，以便于对更复杂的问题抽象建模、定量分析。这是下一章要讲解的内容。

4

Discrete Random Variables

离散随机变量

取值为有限个或可数无穷个，对应概率质量函数 PMF



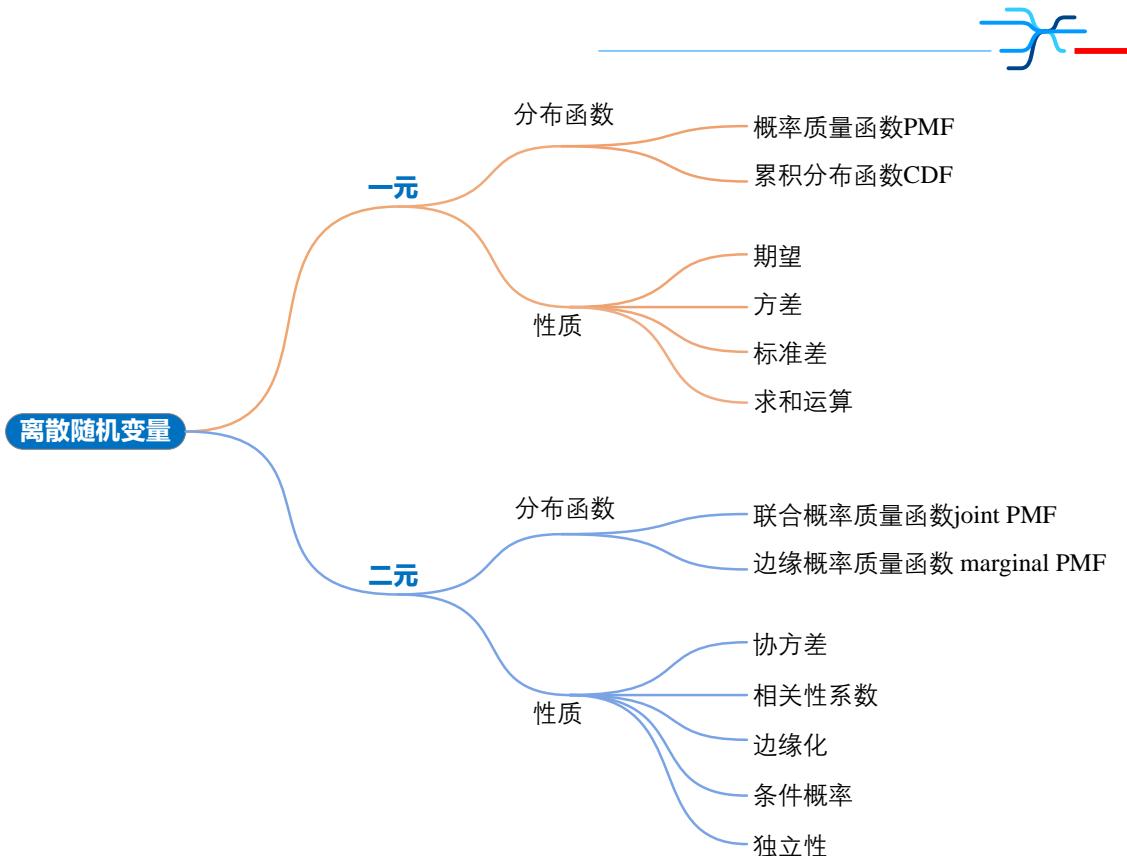
我，一个无数原子组成的宇宙，又是整个宇宙的一粒原子。

I, a universe of atoms, an atom in the universe.

——理查德·费曼 (Richard P. Feynman) | 美国理论物理学家 | 1918 ~ 1988



- ◀ `numpy.sort()` 排序
- ◀ `seaborn.heatmap()` 产生热图
- ◀ `seaborn.histplot()` 绘制频数/概率/概率密度直方图
- ◀ `seaborn.scatterplot()` 绘制散点图



4.1 随机：天地不仁，以万物为刍狗

随机试验

在一定条件下，出现的可能结果不止一个，事前无法确切知道哪一个结果一定会出现，但大量重复试验中结果具有统计规律的现象称为随机现象。

随机试验 (random experiment) 是指在相同条件下对某个随机现象进行的大量重复观测。随机试验需要满足如下条件：

- ▶ 可重复，在相同条件下试验可以重复进行；
- ▶ 样本空间明确，每次试验的可能结果不止一个，并且能事先明确试验的所有可能结果；
- ▶ 单次试验结果不确定，进行一次试验之前不能确定哪一个结果会出现，但必然出现样本空间中的一个。

简单来说，随机试验是指在相同的条件下，每次实验可能出现结果不确定，但是可以用概率来描述可能的结果。例如，投硬币、掷色子等就是随机试验。

两种随机变量：离散、连续

随机变量 (random variable) 是指在一次试验中可能出现不同取值的量，其取值由随机事件的结果决定。随机变量可以看做一个函数，它将样本数值赋给试验结果。换句话说，它是试验样本空间到实数集合的函数。比如上一章为了方便表达“抛三枚色子试验”中三枚色子各自点数，我们定义了 X_1 、 X_2 、 X_3 ，它们都是随机变量。

随机变量分为两种——**离散** (discrete)、**连续** (continuous)。

如果随机变量的所有取值能够一一列举出来，可以是有限个或可数无穷个，这种随机变量被称作**离散随机变量** (discrete random variable)。

比如，投一枚硬币结果正面为 1、反面为 0。掷一枚色子得到的点数为 1、2、3、4、5、6 中的一个值。再比如，鸢尾花的标签有三种——setosa (C_1)、versicolour (C_2)、virginica (C_3)。上一章介绍的古典概率针对离散型随机变量。

与之相对的是，**连续随机变量** (continuous random variable)。连续随机变量取值可能对应全部实数，或者数轴上某一区间。比如，温度、人的身高体重都是连续随机变量。再比如，鸢尾花花萼长度、花萼宽度、花瓣长度、花瓣宽度也都可以视作连续随机变量。

字母

本书用大写斜体字母表达随机变量，比如 X 、 Y 、 Z 、 X_1 、 X_2 、 Y_1 、 Y_2 等。

用小写字母表达随机变量取值，比如 x 、 y 、 x_1 、 x_2 、 y_1 、 y_2 、 i 、 j 、 k 等。其中， x 、 y 、 x_1 、 x_2 、 y_1 、 y_2 等通用于离散、连续随机变量，而序号 i 、 j 、 k 一般用于离散随机变量。

简单来说， X 、 Y 、 Z 、 X_1 、 X_2 、 Y_1 、 Y_2 等替代描述随机试验结果的描述性文字。而 x 、 y 、 x_1 、 x_2 等相当于函数的输入变量，它们主要用在 **概率密度函数** (probability density function, PDF)、**概率质量函数** (probability mass function, PMF) 中。

如图 1 所示，抛一枚色子试验中，令随机变量 X 为色子点数， $X = x$ ， x 代表取值。也就是说， X 的取值为变量 x 。举个例子， $\Pr(X = x)$ 为事件 $\{X = x\}$ 的概率， x 表示随机变量 X 的取值。当然我们可以把数值直接赋值给随机变量，比如 $\Pr(X = 5)$ 。

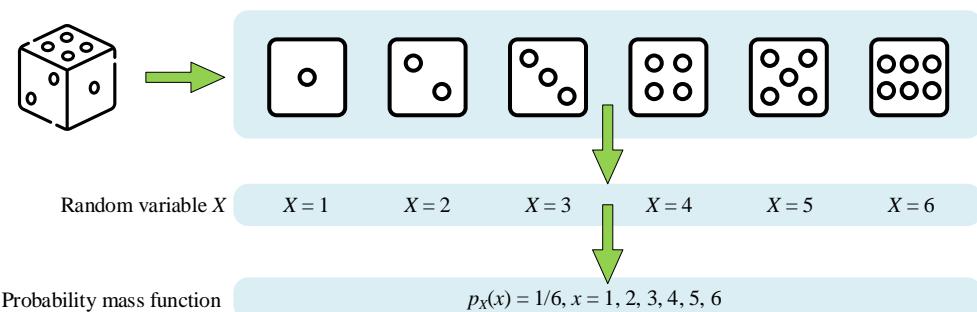


图 1. 随机试验、随机变量、概率质量函数三者关系

两种概率分布函数

研究随机变量取值的统计规律是概率论重要目的之一。概率分布函数是对统计规律的简化和抽象。图 2 比较两种概率分布函数——概率质量函数 PMF、概率密度函数 PDF。

白话来说，概率质量函数 PMF、概率密度函数 PDF 就是两种对样本空间概率为 1“切片、切块”、“切丝、切条”的不同方法。本章后续还会沿着这个思路继续讨论。

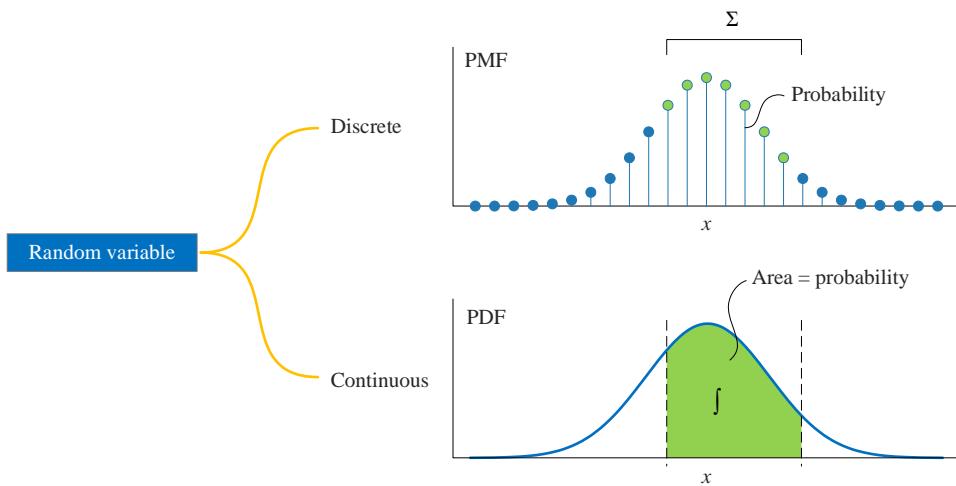


图 2. 比较概率质量函数、概率密度函数

概率质量函数 PMF

图 2 上图所示，**概率质量函数** (probability mass function, PMF) 是离散随机变量在特定取值上的概率。

⚠ 注意，很多教材翻译把 PMF 翻译做“分布列”，本书则将其直译为概率质量函数。

概率质量函数本质上就是概率，因此本书很多时候也直接称之为概率。此外，本书大多时候将概率质量函数直接简写为 PMF。

本书用小字斜体字母 p 表达 PMF，比如随机变量 X 的概率质量函数记做 $p_X(x)$ 。下角标 x 代表描述随机试验的随机变量，概率质量函数的输入为变量 x 。而概率质量函数 $p_X(x)$ 的输出则为“概率值”。

和函数一样，概率质量函数的输入也可以不止一个。比如， $p_{X,Y}(x, y)$ 代表 (X, Y) 的联合概率质量函数。 $p_{X,Y}(x, y)$ 的输入为 (x, y) ，函数的输出为“概率值”。本章后文将专门以二元、三元概率质量函数为例讲解多元概率质量函数。

$p_X(x)$ 本身就是“概率值”，因此计算离散随机变量 X 取不同值时的概率，我们使用求和运算。因此， $p_X(x)$ 对应的数学运算符是 Σ 。

⚠ 注意，有些资料为了方便，将 $p_X(x)$ 简写为 $p(x)$ ， $p_{X,Y}(x, y)$ 简做 $p(x, y)$ 。

抛一枚硬币

举一个例子，抛一枚硬币试验中，令 X_1 为正面朝上数量， X_1 的样本空间为 $\{0, 1\}$ 。 $X_1 = 1$ 代表硬币正面朝上， $X_1 = 0$ 代表硬币反面朝上。

随机变量 X_1 的 PMF 为：

$$p_{X_1}(x_1) = \begin{cases} 1/2 & x_1 = 0 \\ 1/2 & x_1 = 1 \end{cases} \quad (1)$$

相信读者已经对图 3 不陌生，我们在图像上增加标注，水平轴加 x_1 代表 PMF 输入，纵轴改为 PMF， $p_{X_1}(x_1)$ 代表概率质量函数。

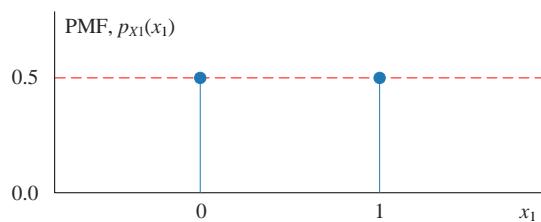


图 3. 随机变量 X_1 的 PMF

如果同时定义 X_2 为反面朝上的数量， X_2 的样本空间也是 $\{0, 1\}$ 。 $X_2 = 1$ 代表硬币反面朝上， $X_2 = 0$ 代表硬币反面朝下。

X_2 的 PMF 为：

$$p_{X_2}(x_2) = \begin{cases} 1/2 & x_2 = 0 \\ 1/2 & x_2 = 1 \end{cases} \quad (2)$$

显然，随机变量 X_1 和 X_2 的关系为 $X_1 + X_2 = 1$ ，具体如图 4 所示。显然 X_1 和 X_2 不独立，大家很快就会发现这种量化关系叫做负相关。

读到这里大家可能已经意识到，在概率质量函数中引入下角标 X_1 和 X_2 能帮助我们区分 $p_{X_1}(x_1)$ 、 $p_{X_2}(x_2)$ 这两个不同的 PMF。

⚠ 注意，本书中随机变量和变量形式上对应，比如 $p_{X_1}(x_1)$ 、 $p_{X_2}(x_2)$ 、 $p_X(x)$ 、 $p_Y(y)$ 。

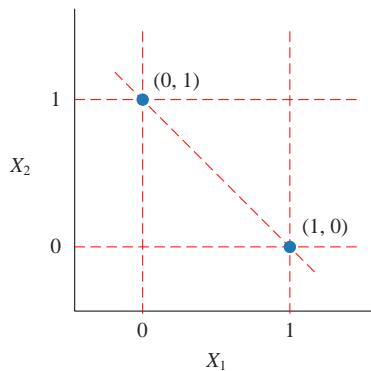


图 4. X_1 和 X_2 的量化关系

抛一个色子

再举一个例子，抛一枚色子试验，令离散随机变量 X 为色子点数。如图 5 所示， X 的 PMF 为：

$$p_X(x) = \begin{cases} 1/6 & x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

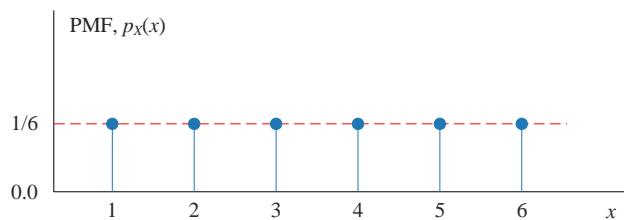


图 5. 离散随机变量 X 的 PMF

随机变量的函数

X 为一个随机变量，对 X 进行函数变换，可以得到其他的随机变量 Y ：

$$Y = h(X) \quad (4)$$

特别地，如果 $h()$ 为线性函数，从 X 到 Y 进行的是线性变换，比如：

$$Y = h(X) = aX + b \quad (5)$$

举个例子，本书前文在抛一枚硬币试验中，令随机变量 X_1 为获得正面的数量，即获得正面时结果为 1，反面结果为 0。

如果，设定一个随机变量 Y ，在硬币为正面时 $Y=1$ ，但是反面时 $Y=-1$ 。那么 X_1 和 Y 的关系如下：

$$Y = 2X_1 - 1 \quad (6)$$



本书第 14 章将专门介绍随机变量的线性变换。

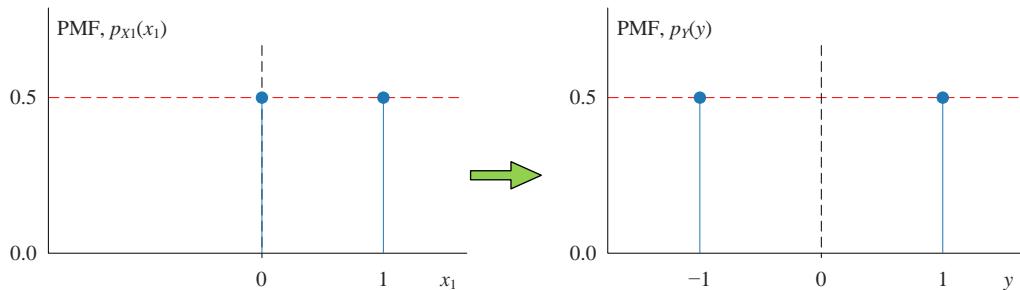


图 6. 随机变量 X_1 线性变换得到 Y 的过程

抛两个色子

上一章讲过一个例子，一次抛两个色子，第一个色子点数设为 X_1 ，第二枚色子的点数为 X_2 。 X_1 和 X_2 可以进行各种数学运算获得随机变量 Y 。

Y 本身有自己的样本空间，样本空间的每个样本都对应特定概率值。利用本章前文内容，我们可以把 $Y=y$ 的概率值写成概率质量函数 $p_Y(y)$ 。

表 1 总结各种“花式玩法”样本空间，以及概率质量函数 $p_Y(y)$ 。表 1 中概率质量函数图像的横纵轴取值范围完全相同。请大家逐个分析，特别注意概率质量函数的分布规律。

表 1. 基于抛两枚色子试验结果的更多花式玩法

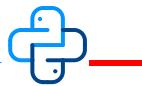
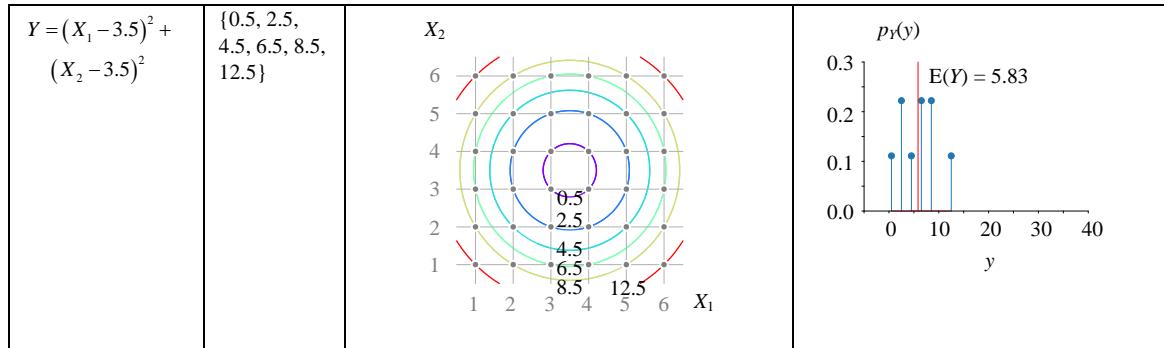
随机变量的函数	样本空间	样本位置	概率质量函数
---------	------	------	--------

$Y = X_1$	$\{1, 2, 3, 4, 5, 6\}$	<table border="1"> <thead> <tr> <th>$X_2 \backslash X_1$</th> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th>6</th> </tr> </thead> <tbody> <tr> <th>1</th> <td>1</td> <td>2</td> <td>3</td> <td>4</td> <td>5</td> <td>6</td> </tr> <tr> <th>2</th> <td>2</td> <td>3</td> <td>4</td> <td>5</td> <td>6</td> <td></td> </tr> <tr> <th>3</th> <td>3</td> <td>4</td> <td>5</td> <td>6</td> <td></td> <td></td> </tr> <tr> <th>4</th> <td>4</td> <td>5</td> <td>6</td> <td></td> <td></td> <td></td> </tr> <tr> <th>5</th> <td>5</td> <td>6</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <th>6</th> <td>6</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>	$X_2 \backslash X_1$	1	2	3	4	5	6	1	1	2	3	4	5	6	2	2	3	4	5	6		3	3	4	5	6			4	4	5	6				5	5	6					6	6						
$X_2 \backslash X_1$	1	2	3	4	5	6																																														
1	1	2	3	4	5	6																																														
2	2	3	4	5	6																																															
3	3	4	5	6																																																
4	4	5	6																																																	
5	5	6																																																		
6	6																																																			
$Y = X_1^2$	$\{1, 4, 9, 16, 25, 36\}$	<table border="1"> <thead> <tr> <th>$X_2 \backslash X_1$</th> <th>1</th> <th>4</th> <th>9</th> <th>16</th> <th>25</th> <th>36</th> </tr> </thead> <tbody> <tr> <th>1</th> <td>1</td> <td>4</td> <td>9</td> <td>16</td> <td>25</td> <td>36</td> </tr> <tr> <th>2</th> <td>2</td> <td>3</td> <td>4</td> <td>5</td> <td>6</td> <td></td> </tr> <tr> <th>3</th> <td>3</td> <td>4</td> <td>5</td> <td>6</td> <td></td> <td></td> </tr> <tr> <th>4</th> <td>4</td> <td>5</td> <td>6</td> <td></td> <td></td> <td></td> </tr> <tr> <th>5</th> <td>5</td> <td>6</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <th>6</th> <td>6</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>	$X_2 \backslash X_1$	1	4	9	16	25	36	1	1	4	9	16	25	36	2	2	3	4	5	6		3	3	4	5	6			4	4	5	6				5	5	6					6	6						
$X_2 \backslash X_1$	1	4	9	16	25	36																																														
1	1	4	9	16	25	36																																														
2	2	3	4	5	6																																															
3	3	4	5	6																																																
4	4	5	6																																																	
5	5	6																																																		
6	6																																																			
$Y = X_1 + X_2$	$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$	<table border="1"> <thead> <tr> <th>$X_2 \backslash X_1$</th> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th>6</th> </tr> </thead> <tbody> <tr> <th>1</th> <td>2</td> <td>3</td> <td>4</td> <td>5</td> <td>6</td> <td></td> </tr> <tr> <th>2</th> <td>3</td> <td>4</td> <td>5</td> <td>6</td> <td></td> <td></td> </tr> <tr> <th>3</th> <td>4</td> <td>5</td> <td>6</td> <td></td> <td></td> <td></td> </tr> <tr> <th>4</th> <td>5</td> <td>6</td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <th>5</th> <td>6</td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> <tr> <th>6</th> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td> </tr> </tbody> </table>	$X_2 \backslash X_1$	1	2	3	4	5	6	1	2	3	4	5	6		2	3	4	5	6			3	4	5	6				4	5	6					5	6						6							
$X_2 \backslash X_1$	1	2	3	4	5	6																																														
1	2	3	4	5	6																																															
2	3	4	5	6																																																
3	4	5	6																																																	
4	5	6																																																		
5	6																																																			
6																																																				
$Y = \frac{X_1 + X_2}{2}$	$\{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0\}$	<table border="1"> <thead> <tr> <th>$X_2 \backslash X_1$</th> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th>6</th> </tr> </thead> <tbody> <tr> <th>1</th> <td>1.0</td> <td>1.5</td> <td>2.0</td> <td>2.5</td> <td>3.0</td> <td></td> </tr> <tr> <th>2</th> <td>1.5</td> <td>2.0</td> <td>2.5</td> <td>3.0</td> <td>3.5</td> <td></td> </tr> <tr> <th>3</th> <td>2.0</td> <td>2.5</td> <td>3.0</td> <td>3.5</td> <td>4.0</td> <td></td> </tr> <tr> <th>4</th> <td>2.5</td> <td>3.0</td> <td>3.5</td> <td>4.0</td> <td>4.5</td> <td></td> </tr> <tr> <th>5</th> <td>3.0</td> <td>3.5</td> <td>4.0</td> <td>4.5</td> <td>5.0</td> <td></td> </tr> <tr> <th>6</th> <td>3.5</td> <td>4.0</td> <td>4.5</td> <td>5.0</td> <td>5.5</td> <td>6.0</td> </tr> </tbody> </table>	$X_2 \backslash X_1$	1	2	3	4	5	6	1	1.0	1.5	2.0	2.5	3.0		2	1.5	2.0	2.5	3.0	3.5		3	2.0	2.5	3.0	3.5	4.0		4	2.5	3.0	3.5	4.0	4.5		5	3.0	3.5	4.0	4.5	5.0		6	3.5	4.0	4.5	5.0	5.5	6.0	
$X_2 \backslash X_1$	1	2	3	4	5	6																																														
1	1.0	1.5	2.0	2.5	3.0																																															
2	1.5	2.0	2.5	3.0	3.5																																															
3	2.0	2.5	3.0	3.5	4.0																																															
4	2.5	3.0	3.5	4.0	4.5																																															
5	3.0	3.5	4.0	4.5	5.0																																															
6	3.5	4.0	4.5	5.0	5.5	6.0																																														
$Y = \frac{X_1 + X_2 - 7}{2}$	$\{-2.5, -2.0, -1.5, -1.0, -0.5, 0.0, 0.5, 1.0, 1.5, 2.0, 2.5\}$	<table border="1"> <thead> <tr> <th>$X_2 \backslash X_1$</th> <th>1</th> <th>2</th> <th>3</th> <th>4</th> <th>5</th> <th>6</th> </tr> </thead> <tbody> <tr> <th>1</th> <td>-2.5</td> <td>-2.0</td> <td>-1.5</td> <td>-1.0</td> <td>-0.5</td> <td>0.0</td> </tr> <tr> <th>2</th> <td>-2.0</td> <td>-1.5</td> <td>-1.0</td> <td>-0.5</td> <td>0.0</td> <td>0.5</td> </tr> <tr> <th>3</th> <td>-1.5</td> <td>-1.0</td> <td>-0.5</td> <td>0.0</td> <td>0.5</td> <td>1.0</td> </tr> <tr> <th>4</th> <td>-1.0</td> <td>-0.5</td> <td>0.0</td> <td>0.5</td> <td>1.0</td> <td>1.5</td> </tr> <tr> <th>5</th> <td>-0.5</td> <td>0.0</td> <td>0.5</td> <td>1.0</td> <td>1.5</td> <td>2.0</td> </tr> <tr> <th>6</th> <td>0.0</td> <td>0.5</td> <td>1.0</td> <td>1.5</td> <td>2.0</td> <td>2.5</td> </tr> </tbody> </table>	$X_2 \backslash X_1$	1	2	3	4	5	6	1	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	2	-2.0	-1.5	-1.0	-0.5	0.0	0.5	3	-1.5	-1.0	-0.5	0.0	0.5	1.0	4	-1.0	-0.5	0.0	0.5	1.0	1.5	5	-0.5	0.0	0.5	1.0	1.5	2.0	6	0.0	0.5	1.0	1.5	2.0	2.5	
$X_2 \backslash X_1$	1	2	3	4	5	6																																														
1	-2.5	-2.0	-1.5	-1.0	-0.5	0.0																																														
2	-2.0	-1.5	-1.0	-0.5	0.0	0.5																																														
3	-1.5	-1.0	-0.5	0.0	0.5	1.0																																														
4	-1.0	-0.5	0.0	0.5	1.0	1.5																																														
5	-0.5	0.0	0.5	1.0	1.5	2.0																																														
6	0.0	0.5	1.0	1.5	2.0	2.5																																														

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>
本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$Y = X_1 X_2$	$\{1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 15, 16, 18, 20, 24, 25, 30, 36\}$		
$Y = \frac{X_1}{X_2}$	$\{0.166, 0.2, 0.25, 0.333, 0.4, 0.5, 0.6, 0.666, 0.75, 0.8, 0.833, 1.0, 1.2, 1.25, 1.333, 1.5, 1.666, 2.0, 2.5, 3.0, 4.0, 5.0, 6.0\}$		
$Y = X_1 - X_2$	$\{-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5\}$		
$Y = X_1 - X_2 $	$\{0, 1, 2, 3, 4, 5\}$		



代码 Bk5_Ch04_01.py 绘制表 1 中图像。学完本章后续内容后，请大家修改代码计算 Y 标准差 $\text{std}(Y)$ ，并在火柴梗图上展示 $E(Y) \pm \text{std}(Y)$ 。

归一律

一元离散随机变量 X 的概率质量函数 $p_X(x)$ 有如下重要性质：

$$\sum_x p_X(x) = 1, \quad 0 \leq p_X(x) \leq 1 \quad \begin{array}{c} \text{dots} \\ \vdots \end{array} \quad (7)$$

上式实际上就是“穷举法”，即遍历所有 X 取值，将它们的概率值求和，结果为 1。“穷举法”也叫归一律。

⚠ 值得强调的是，概率质量函数 $p_X(x)$ 最大取值为 1。

概率密度函数 PDF

与 PMF 相对的是**概率密度函数** (probability density function, PDF)。PDF 对应连续随机变量，本书用小写斜体字母 f 表达 PDF，比如连续随机变量 X 的概率密度函数记做 $f_X(x)$ 。

⚠ 注意，在本书第 20、21 章中讲解贝叶斯推断时，为了方便，概率质量函数、概率密度函数都用 $f()$ 。

当连续随机变量取不同值时，概率密度函数 $f_X(x)$ 用积分方式得到概率值。因此， $f_X(x)$ 对应的数学运算符是积分符号 \int 。

举个例子，连续随机变量 X 服从标准正态分布 $N(0, 1)$ ，其 PDF 为：

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad (8)$$

其中，变量 x 的取值范围为整个实数轴；对于标准正态分布 $N(0, 1)$ ，其 $f_X(x)$ 取值可以无限接近 0，却不为 0。

当 $x = 0$ 时， $f_X(x)$ 约为 0.4，这个值是概率密度，不是概率。只有对连续随机变量 PDF 在指定区间内进行积分后结果才“可能”是概率。

⚠ 注意，联合概率密度函数 $f_{X_1, X_2, X_3}(x_1, x_2, x_3)$ “偏积分”结果还是概率密度。 $f_{X_1, X_2, X_3}(x_1, x_2, x_3)$ 三重积分结果才是概率值。

⚠ 值得反复强调的是，PMF 本身就是概率，对应的数学工具为 Σ 求和。PDF 积分后才可能是概率，对应的数学工具为 \int 积分。

一元连续随机变量 X 的概率密度函数 $f_X(x)$ 也有如下重要性质：

$$\int_{-\infty}^{+\infty} f_X(x) dx = 1, \quad f_X(x) \geq 0 \quad \text{Area} = 1 \quad (9)$$

上式也相当于是“穷举法”。

⚠ 注意，概率密度函数 $f_X(x)$ 取值非负，但是不要求小于 1。本书后续将给出具体示例。

概率质量函数 PMF、概率密度函数 PDF 是特殊的函数。特殊之处在于它们的输入为随机变量的取值，输出为概率质量、概率密度。但是，本质上，它们又都是函数。所以，我们可以把函数的分析工具用在概率质量函数 PMF、概率密度函数 PDF 上。

→ 本章和下一章首先讲解离散随机变量、离散分布。本书第 6、7 章讲解连续随机变量、连续分布。

区分符号

有必要再次区分本系列丛书的容易混淆的代数、线性代数、概率统计符号。

→ 以下内容主要来自《矩阵力量》第 23 章，稍作改动。

粗体、斜体、小写 x 为列向量。从概率统计的角度， x 可以代表随机变量 X 采样得到的样本数据，偶尔也代表 X 总体样本。随机变量 X 样本“无序”集合为 $X = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ 。很多时候，随机变量 X 样本本身也可以看成“有序”的数组，即向量。

粗体、斜体、小写、加下标序号的 x_i 为列向量，下角标仅仅是序号，以便区分，比如 x_1 、 x_2 、 x_j 、 x_D 等等。从概率统计的角度， x_i 可以代表随机变量 X_i 样本数据，也可以表达 X_i 总体数据。

行向量 $x^{(i)}$ 代表一个具有多个特征的样本点。

⚠ 注意，在机器学习算法中，为了方便， $x^{(i)}$ 偶尔也代表列向量。

从代数角度，斜体、小写、非粗体 x_1 代表变量，下角标代表变量序号。这种记法常用在函数解析式中，比如线性回归解析式 $y = x_1 + x_2$ 。在概率质量函数、概率密度函数中，它们也用做 PMF、PDF 函数输入，比如 $p_{X1}(x_1)$ 、 $f_{X2}(x_2)$ 。

\mathbf{x} 也代表变量构成的列向量， $\mathbf{x} = [x_1, x_2, \dots, x_D]^T$ ，比如多元概率密度函数 $f_X(\mathbf{x})$ 的输入。

$x^{(1)}$ 代表变量 x 的一个取值，或代表随机变量 X 的一个取值。

而 $x_1^{(1)}$ 代表变量 x_1 的一个取值，或代表随机变量 X_1 的一个取值，比如 $X_1 = \{x_1^{(1)}, x_1^{(2)}, \dots, x_1^{(n)}\}$ 。

粗体、斜体、大写 X 则专门用来表达多行、多列的数据矩阵， $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D]$ 。数据矩阵 X 中第 i 行、第 j 列元素则记做 x_{ij} 。

多元线性回归中， \mathbf{X} 也叫**设计矩阵** (design matrix)。设计矩阵第一列一般有全 1 列向量。

我们还会用粗体、斜体、小写希腊字母 χ (chi, 读作/kai/) 代表 D 维随机变量构成的列向量， $\chi = [X_1, X_2, \dots, X_D]^T$ 。希腊字母 χ 主要用在多元概率统计中，比如，多元概率密度函数 $f_\chi(\mathbf{x})$ 、期望值列向量 $E(\chi)$ 。

4.2 期望值：随机变量的可能取值加权平均

期望值

离散随机变量 X 有 n 个取值 $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ ， X 的**期望** (expectation)，也叫**期望值** (expected value)， $E(X)$ 为：

$$E(X) = \mu_X = x^{(1)} p_X(x^{(1)}) + x^{(2)} p_X(x^{(2)}) + \dots + x^{(n)} p_X(x^{(n)}) = \sum_{i=1}^n x^{(i)} \cdot \underbrace{p_X(x^{(i)})}_{\text{Weight}} \quad (10)$$

上式相当于加权平均数，边缘 PMF $p_X(x)$ 代表权重。

运算符 $E()$ 把随机变量一系列取值转化成了一个标量数值，这相当于降维。如图 7 所示，从矩阵乘法角度，计算期望值 $E(X)$ 相当于将 X 这个维度折叠。

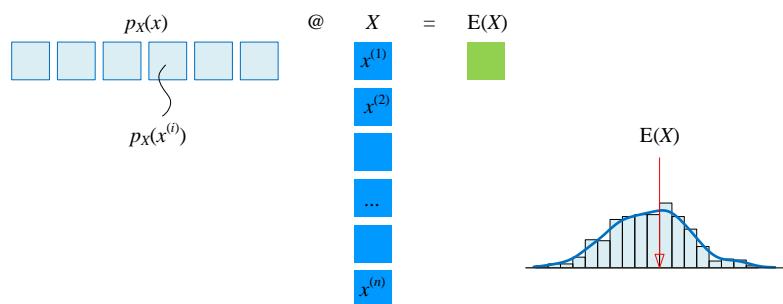
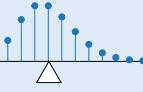


图 7. 计算离散随机变量 X 期望值/均值

为了方便，我们经常把(10)简写作：

$$\mathbb{E}(X) = \sum_x x \cdot p_X(x)$$

(11)

$\sum_x (\cdot)$ 代表对 x 的遍历求和，也就是穷举。求加权平均值时，权重之和为 1，也就是说边缘 PMF $p_X(x)$ 满足 $\sum_x p_X(x) = 1$ 。特别是对于多元随机变量，我们也经常把期望值（均值）叫做质心（centroid）。

举个例子

图 5 中随机变量 X 的期望值为：

$$\mathbb{E}(X) = \sum_x x \cdot \underbrace{p_X(x)}_{\text{Weight}} = \sum_x x \cdot \underbrace{\frac{1}{6}}_{\text{Weight}} = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5 \quad (12)$$

大家已经发现上式中随机变量 X 的概率质量函数为定值。这和求样本均值的情况类似。求 n 个样本均值时，每个样本赋予的权重为 $1/n$ ，即每个样本权重相同。

图 8 所示为投色子试验均值随试验次数变化。随着重复次数接近无穷大，试验结果的算术平均值（试验概率）收敛于 3.5（理论值）。

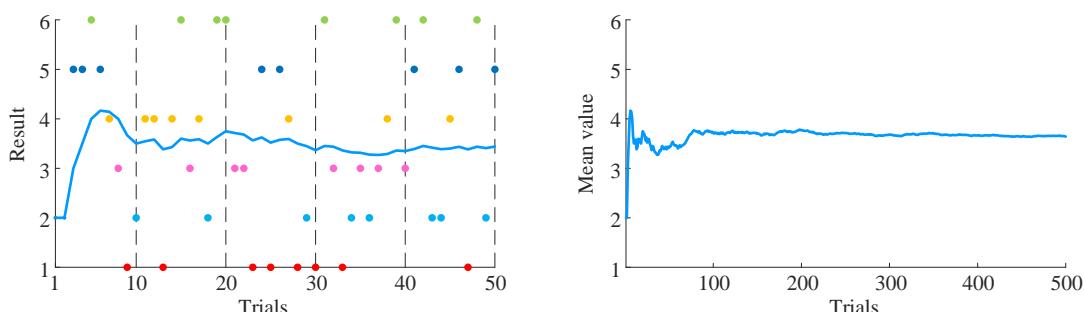


图 8. 投色子试验均值随试验次数变化

重要性质

请大家注意以下几个有关期望的性质：

$$\begin{aligned} \mathbb{E}(aX) &= a\mathbb{E}(X) \\ \mathbb{E}(X+Y) &= \mathbb{E}(X)+\mathbb{E}(Y) \end{aligned} \quad (13)$$

如果 X 和 Y 独立：

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \quad (14)$$

此外，请大家注意：

$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i \mathbb{E}(X_i) \quad (15)$$

特别地，当 $n = 2$ 时，上式可以写成：

$$\mathbb{E}(a_1 X_1 + a_2 X_2) = a_1 \mathbb{E}(X_1) + a_2 \mathbb{E}(X_2) \quad (16)$$

(16) 可以写成如下矩阵乘法运算：

$$\mathbb{E}(a_1 X_1 + a_2 X_2) = [a_1 \quad a_2] \underbrace{\begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{bmatrix}}_{\mu} \quad (17)$$

同理，(15) 可以写成：

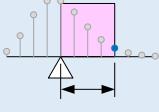
$$\mathbb{E}\left(\sum_{i=1}^n a_i X_i\right) = [a_1 \quad a_2 \quad \cdots \quad a_n] \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_n) \end{bmatrix} \quad (18)$$

请大家自己把矩阵乘法运算示意图画出来。

4.2 方差：随机变量离期望距离平方的平均值

方差

随机变量 X 另外一个重要特征是**方差** (variance)，记做 $\text{var}(X)$ 。对于离散随机变量 X ，方差用来度量 X 和数学期望 $\mathbb{E}(X)$ 之间的偏离程度。具体定义为：

$$\text{var}(X) = \mathbb{E}\left[\underbrace{\left(X - \mathbb{E}(X)\right)^2}_{\text{Deviation}}\right] = \sum_x \underbrace{\left(x - \mathbb{E}(X)\right)^2}_{\text{Demean}} \cdot \underbrace{p_X(x)}_{\text{Weight}} \quad (19)$$


上式中 $x - \mathbb{E}(X)$ 代表以期望值 $\mathbb{E}(X)$ 为参照，样本点 x 的偏离量。

如图 9 所示， $X - \mathbb{E}(X)$ 代表去均值 (demean)，也叫**中心化** (centralize)。

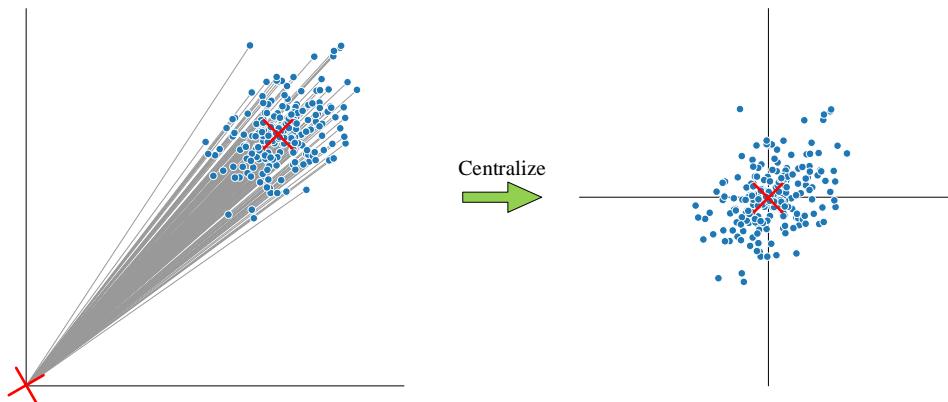


图 9. 样本去均值

观察 (19)，容易发现方差实际上是 $(X - E(X))^2$ 的期望值。(19) 就是求 $(x - E(X))^2$ 的加权平均数，权重为 $p_X(x)$ 。从几何角度， $(X - E(X))^2$ 代表以 $|X - E(X)|$ 为边长的正方形的面积。而对于离散随机变量， $p_X(x)$ 就是权重，体现不同样本重要性。

举个例子

图 5 对应的方差为：

$$\begin{aligned} \text{var}(X) &= \frac{1}{6} \times (1 - 3.5)^2 + \frac{1}{6} \times (2 - 3.5)^2 + \frac{1}{6} \times (3 - 3.5)^2 + \frac{1}{6} \times (4 - 3.5)^2 + \frac{1}{6} \times (5 - 3.5)^2 + \frac{1}{6} \times (6 - 3.5)^2 \\ &= \frac{1}{6} \times \left(\frac{25}{4} + \frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4} + \frac{25}{4} \right) = \frac{35}{12} \approx 2.9167 \end{aligned} \quad (20)$$

注意，本书前文在计算样本方差时，分母除以 $n - 1$ 。而 (20) 分母相当于除以 n ，这是因为 (20) 是对总体样本求方差。而且，恰好 X 取 $1 \sim 6$ 这六个不同值是对应的概率相等。

也就是说，当离散随机变量 X 等概率时，概率质量函数为：

$$p_X(x) = \frac{1}{n} \quad (21)$$

(19) 可以写成：

$$\text{var}(X) = \frac{1}{n} \sum_x (x - E(X))^2 \quad (22)$$

再次强调，上式是求离散随机变量方差的一种特殊情况（离散均匀分布）。统计中，样本的方差计算方法类似上式，不过要将分母中的 n 换成 $n - 1$ 。

技巧：方差计算

方差有个简便算法：

$$\text{var}(X) = \underbrace{\mathbb{E}(X^2)}_{\text{Expectation of } X^2} - \underbrace{\mathbb{E}(X)^2}_{\text{Square of } \mathbb{E}(X)} \quad (23)$$

其中， $\mathbb{E}(X^2)$ 为：

$$\underbrace{\mathbb{E}(X^2)}_{\text{Expectation of } X^2} = \sum_x x^2 \cdot \underbrace{p_X(x)}_{\text{Weight}} \quad (24)$$

(23) 的推导过程如下所示：

$$\begin{aligned} \text{var}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2 - 2X \cdot \mathbb{E}(X) + \mathbb{E}(X)^2) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X) \cdot \mathbb{E}(X) + \mathbb{E}(X)^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \end{aligned} \quad (25)$$

注意，(23) 也适用于连续随机变量。

请大家尝试使用 (25) 计算 (20) 的方差。

几何意义

下面我们聊聊 (25) 的几何含义。

方差度量离散程度，本质上来说是“自己”和“自己”比较的产物。前一个“自己”是 X 每个样本，后一个“自己”是代表 X 整体位置的期望值 $\mathbb{E}(X)$ 。

如图 10 所示，方差 $\text{var}(X)$ 代表样本以质心 (centroid) 为基准的离散程度。

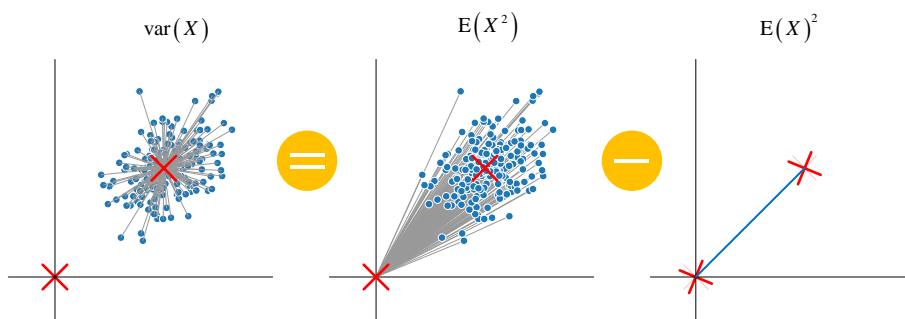


图 10. 几何视角理解计算方差技巧

(23) 中，计算方差 $\text{var}(X)$ 有 $\mathbb{E}(X^2)$ 和 $-\mathbb{E}(X)^2$ 两部分。

$\mathbb{E}(X^2)$ 度量 X 样本以原点 (origin) 为基准的离散程度。

$\mathbb{E}(X)^2$ 则代表 X 整体，即 $\mathbb{E}(X)$ ，相对于原点的离散程度。 $-\mathbb{E}(X)^2$ 中的“负号”代表将基准从原点移到质心。

特别地，当 X 的质心位于原点，即 $E(X) = 0$ 时， $\text{var}(X)$ 为：

$$\text{var}(X) = E(X^2) \quad (26)$$

标准差

标准差 (standard deviation) 是方差的平方根：

$$\text{std}(X) = \sigma_X = \sqrt{\text{var}(X)} \quad (27)$$

方差既然可以用来度量“离散程度”，为什么我们还需要标准差？

简单来说，标准差 σ_X 、期望值 $E(X)$ 、随机变量 X 为同一量纲。比如，鸢尾花花萼长度 X 的单位是 cm，期望值 $E(X)$ 的单位也是 cm，而 σ_X 的单位也对应是 cm。但是， $\text{var}(X)$ 的量纲是 cm^2 。

需要注意的性质

请大家注意以下方差性质：

$$\begin{aligned} \text{var}(a) &= 0 \\ \text{var}(X + a) &= \text{var}(X) \\ \text{var}(aX) &= a^2 \text{var}(X) \\ \text{var}(aX + b) &= a^2 \text{var}(X) \\ \text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \end{aligned} \quad (28)$$

其中 $\text{cov}(X, Y)$ 为随机变量 X 和 Y 的协方差，本章后续将专门介绍协方差。

请大家注意以下标准差性质：

$$\begin{aligned} \sigma(a) &= 0 \\ \sigma(X + a) &= \sigma(X) \\ \sigma(bX) &= |b|\sigma(X) \\ \sigma(a + bX) &= |b|\sigma(X) \\ \sigma(X + Y) &= \sqrt{\sigma^2(X) + \sigma^2(Y) + 2\rho(X, Y)\sigma(X)\sigma(Y)} \end{aligned} \quad (29)$$

汇总

折叠、总结、汇总、降维、压扁 … 本章及本书后文会用这些字眼形容期望值、方差、标准差。这是因为，计算期望值、方差、标准差时，我们不再关注随机变量样本具体取值，而是在乎某种方式的**汇总** (aggregation)。

期望值、方差、标准差将“数组”转化成特定标量值。因此，这个特定维度相当于被折叠、总结、汇总、降维、压扁 … 对于多元随机变量，我们可以选择某个、某几个维度上完成汇总计算。

如果汇总的形式为期望，它相当于找到随机变量整体的“位置”。如果汇总的形式为方差、标准差，两者都度量随机变量“离散”程度。

其他常用的汇总形式还包括：**计数** (count)、**求和** (sum)、**四分位** (quartile)、**百分位** (percentile)、**最大值** (maximum)、**最小值** (minimum)、**中位数** (median)、**众数** (mode)、偏度、峰度等等。

4.3 累积分布函数 CDF：累加

对于离散随机变量，**累积分布函数** (Cumulative Distribution Function, CDF) 对应概率质量函数的求和。

对于离散随机变量 X ，累积分布函数 $F_X(x)$ 的定义为：

$$F_X(x) = \Pr(X \leq x) = \sum_{t \leq x} p_X(t) \quad (30)$$

上式相当于累加概念，累加从 X 最小样本值开始并截止于 $X=x$ 。

离散随机变量 X 的取值范围为 $a < X \leq b$ 时，对应的概率可以利用 CDF 计算：

$$\Pr(a < X \leq b) = F_X(b) - F_X(a) \quad (31)$$

图 5 对应的 CDF 图像为图 11。

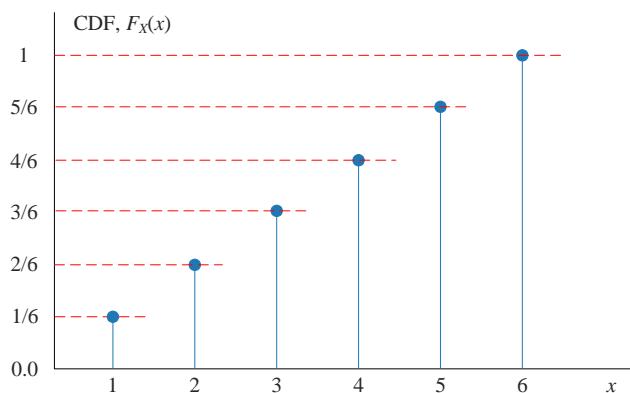


图 11. 随机变量 X 的 CDF

⚠ 注意，对于离散随机变量，区间端点的开闭影响结果。

以图 11 为例，请大家比较以下四个不同开闭区间的概率值：

$$\Pr(1 < X \leq 3) = \frac{1}{3}, \quad \Pr(1 \leq X \leq 3) = \frac{1}{2}, \quad \Pr(1 \leq X < 3) = \frac{1}{3}, \quad \Pr(1 < X < 3) = \frac{1}{6} \quad (32)$$

对于连续随机变量，就没有区间端点的麻烦了。本书第 6 章将展开讲解。

4.4 二元离散随机变量

假设同一个试验中，有两个离散随机变量 X 和 Y 。二元随机变量 (X, Y) 概率取值可以用**联合概率质量函数** (joint Probability Mass Function, joint PMF) $p_{X,Y}(x, y)$ 刻画。

概率质量函数 $p_{X,Y}(x, y)$ 代表事件 $\{X = x, Y = y\}$ 发生的联合概率：

$$\underbrace{p_{X,Y}(x, y)}_{\text{Joint}} = \Pr(X = x, Y = y) = \Pr(X = x \cap Y = y) \quad (33)$$

⚠ 再次强调，对于二元离散随机变量， $p_{X,Y}(x, y)$ 本身就是概率值。

图 12 所示为二元离散随机变量 (X, Y) 的样本空间 Ω ，空间中共有 81 个点。从函数角度来看， $p_{X,Y}(x, y)$ 是一个二元函数。因此，我们可以用二元函数的分析方法来讨论 $p_{X,Y}(x, y)$ 。

→ 《数学要素》第 13 章介绍二元函数，建议大家回顾。

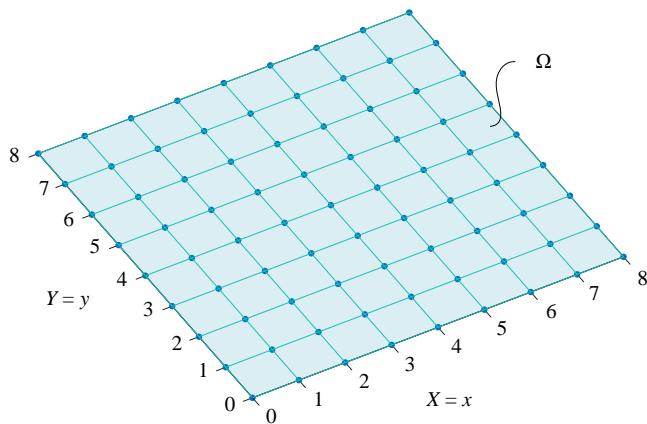


图 12. 二元随机变量的样本空间

取值

图 13 所示为二元联合概率质量函数 $p_{X,Y}(x, y)$ 的取值表格。图 13 同时用热图来可视化 $p_{X,Y}(x, y)$ 。

二元联合概率质量函数 $p_{X,Y}(x, y)$ 也有一条重要的性质：

$$\sum_x \sum_y \underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \sum_y \sum_x \underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = 1, \quad 0 \leq p_{X,Y}(x,y) \leq 1$$

(34)

也就是说，图 13 这幅热图中所有数值（概率，概率质量）求和的结果为 1，和求和顺序无关。

		$X = x$								
		0	1	2	3	4	5	6	7	8
$Y = y$	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	7	0.0000	0.0000	0.0000	0.0001	0.0002	0.0003	0.0004	0.0002	0.0001
	6	0.0000	0.0000	0.0001	0.0005	0.0014	0.0025	0.0030	0.0020	0.0006
	5	0.0000	0.0001	0.0005	0.0022	0.0064	0.0119	0.0138	0.0092	0.0027
	4	0.0000	0.0002	0.0014	0.0064	0.0185	0.0346	0.0404	0.0269	0.0078
	3	0.0000	0.0003	0.0025	0.0119	0.0346	0.0646	0.0753	0.0502	0.0146
	2	0.0000	0.0004	0.0030	0.0138	0.0404	0.0753	0.0879	0.0586	0.0171
	1	0.0000	0.0002	0.0020	0.0092	0.0269	0.0502	0.0586	0.0391	0.0114
	0	0.0000	0.0001	0.0006	0.0027	0.0078	0.0146	0.0171	0.0114	0.0033

图 13. 概率质量函数 $p_{X,Y}(x,y)$ 取值

火柴梗图

二元联合概率质量函数 $p_{X,Y}(x,y)$ 长成什么样子呢？

火柴梗图最适合可视化概率质量函数，如图 14 所示。

⚠ 注意，为了展示火柴梗图分别沿 X 、 Y 方向变化趋势，图 14 将火柴梗散点连线。一般情况，火柴梗图不存在连线。

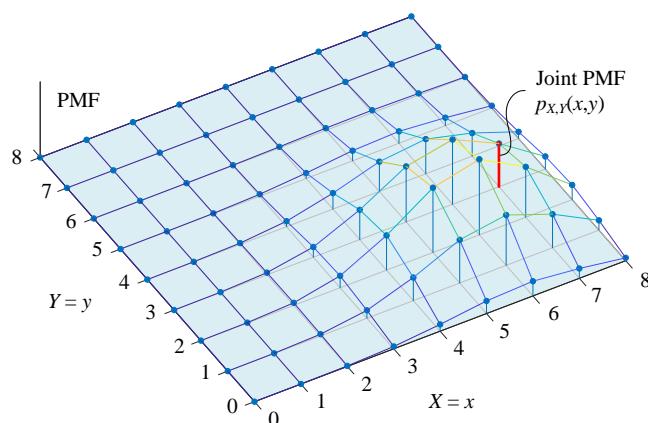


图 14. $p_{X,Y}(x, y)$ 对应的二维火柴梗图

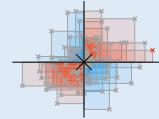
4.5 协方差、相关性系数

本书读者对协方差、相关性系数这两个概念应该不陌生，本节简要介绍如何求解离散随机变量的协方差和相关性系数。

协方差

二元离散随机变量 (X, Y) 的协方差定义为：

$$\text{cov}(X, Y) = E((X - E(X))(Y - E(Y)))$$



(35)

如果 (X, Y) 的概率质量函数为 $p_{X,Y}(x, y)$, X 的取值为 $x^{(i)}$ ($i = 1, 2, \dots, n$), Y 的取值为 $y^{(j)}$ ($j = 1, 2, \dots, m$)。 (35) 可以展开写成：

$$\begin{aligned} \text{cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= \sum_{i=1}^n \sum_{j=1}^m p_{X,Y}(x^{(i)}, y^{(j)}) (x^{(i)} - E(X)) (y^{(j)} - E(Y)) \end{aligned} \quad (36)$$

其中，

$$E(X) = \sum_x x \cdot p_X(x), \quad E(Y) = \sum_y y \cdot p_Y(y) \quad (37)$$

(36) 常简写为：

$$\text{cov}(X, Y) = \sum_x \sum_y p_{X,Y}(x, y) (x - E(X)) (y - E(Y)) \quad (38)$$

类似方差，协方差运算也有如下技巧：

$$\begin{aligned} \text{cov}(X, Y) &= E(XY) - E(X)E(Y) \\ &= \sum_x \sum_y x \cdot y \cdot p_{X,Y}(x, y) - \left(\sum_x x \cdot p_X(x) \right) \cdot \left(\sum_y y \cdot p_Y(y) \right) \end{aligned} \quad (39)$$

(39) 推导过程具体如下：

$$\begin{aligned} \text{cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY - E(X)Y - X E(Y) + E(X E(Y))) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned} \quad (40)$$

建议大家也用类似图 10 的几何视角理解上式。

相关性

(X, Y) 相关性的定义为：

$$\rho_{x,y} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \quad (41)$$

展开得到：

$$\rho_{x,y} = \frac{\text{E}(XY) - \text{E}(X)\text{E}(Y)}{\sqrt{\text{E}(X^2) - \text{E}(X)^2} \sqrt{\text{E}(Y^2) - \text{E}(Y)^2}} \quad (42)$$

相关性的取值范围 $[-1, 1]$ 。相对协方差，相关性更适合横向比较。



本书第 10 章将专门讲解相关性。

协方差性质

请大家注意以下协方差性质：

$$\begin{aligned} \text{cov}(X, a) &= 0 \\ \text{cov}(X, X) &= \text{var}(X) \\ \text{cov}(X, Y) &= \text{cov}(Y, X) \\ \text{cov}(aX, bY) &= ab \text{cov}(X, Y) \\ \text{cov}(X + a, Y + b) &= \text{cov}(X, Y) \\ \text{cov}(aX + bY, Z) &= a \text{cov}(X, Z) + b \text{cov}(Y, Z) \\ \text{cov}(aX + bY, cW + dV) &= ac \text{cov}(X, W) + ad \text{cov}(X, V) + bc \text{cov}(Y, W) + bd \text{cov}(Y, V) \end{aligned} \quad (43)$$

此外，方差和协方差的关系：

$$\text{var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_i a_i^2 \text{var}(X_i) + 2 \sum_{i,j: i < j} a_i a_j \text{cov}(X_i, X_j) = \sum_{i,j} a_i a_j \text{cov}(X_i, X_j) \quad (44)$$

特别地，当 $n = 2$ 时，上式可以写成：

$$\text{var}(a_1 X_1 + a_2 X_2) = a_1^2 \text{var}(X_1) + a_2^2 \text{var}(X_2) + 2a_1 a_2 \text{cov}(X_1, X_2) \quad (45)$$



看到 (45) 大家是否立刻想到我们在《矩阵力量》第 5 章介绍过的二次型 (quadratic form)。

(45) 可以写成如下矩阵乘法运算：

$$\text{var}(a_1 X_1 + a_2 X_2) = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}^T \underbrace{\begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \text{var}(X_2) \end{bmatrix}}_{\Sigma} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} \quad (46)$$

同理，(44) 可以写成：

$$\text{var}\left(\sum_{i=1}^n a_i X_i\right) = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}^T \underbrace{\begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \cdots & \text{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \cdots & \text{cov}(X_n, X_n) \end{bmatrix}}_{\Sigma} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} \quad (47)$$

→ 本书第 14 章将从向量投影视角深入讲解上式。

几何视角

对于如下等式，

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2 \text{cov}(X, Y) \quad (48)$$

即，

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho_{X,Y}\sigma_X\sigma_Y \quad (49)$$

看到上式，大家是否立刻联想到《数学要素》第 3 章介绍的余弦定律 (law of cosines)：

$$c^2 = a^2 + b^2 - 2ab \cos \theta \quad (50)$$

σ_X 、 σ_Y 、 σ_{X+Y} 相当于三角形的三个边， $\rho_{X,Y}$ 相当 σ_X 、 σ_Y 于夹角的余弦值。如图 15 所示，当 $\rho_{X,Y}$ 取不同值时，三角形呈现不同的形态。

→ 另外一个视角就是《矩阵力量》介绍的“标准差向量”，请大家回顾。

特别地，如果 $\rho_{X,Y} = 0$ ，三角形为直角三角形，满足：

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 \quad (51)$$

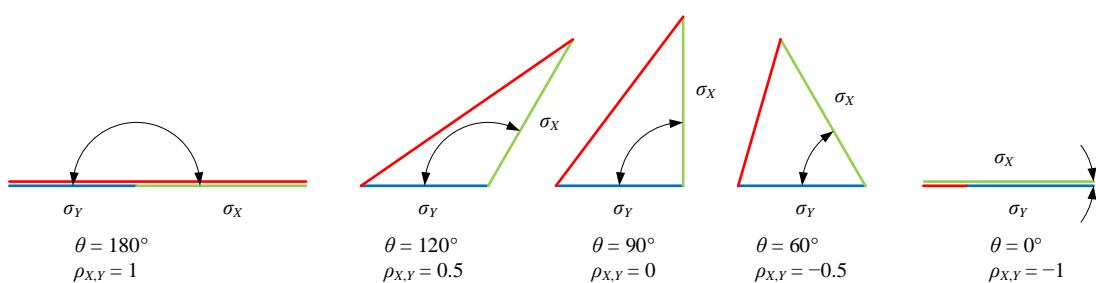


图 15. 将余弦定理用到方差等式



此外，《矩阵力量》第 22 章还专门类比向量内积和协方差，建议大家回顾。

4.6 边缘概率：偏求和，相当于降维

边缘概率 (marginal probability) 是某个事件发生的概率，而与其它事件无关。对于离散随机变量来说，利用全概率定理，也就是穷举法，我们可以把联合概率结果中不需要的那些事件全部合并。合并的过程叫做**边缘化** (marginalization)。

→ 对于多元离散随机变量，边缘化用到的数学工具为《数学要素》第 14 章讲到的“偏求和”。

边缘概率 $p_X(x)$

根据全概率公式，对于二元联合概率质量函数 $p_{X,Y}(x,y)$ ，求解边缘概率 $p_X(x)$ 相当于利用“偏求和”消去 y ：

$$\underbrace{p_X(x)}_{\text{Marginal}} = \sum_{y} \underbrace{p_{X,Y}(x,y)}_{\text{Joint}}$$
(52)

也就是说，在 $X=x$ 取值条件下， $p_{X,Y}(x,y)$ 对所有 y 的求和。

从函数角度来看， $p_{X,Y}(x,y)$ 是个二元函数， $p_X(x)$ 是个一元函数。

从矩阵运算角度来看， $p_{X,Y}(x,y)$ 代表矩阵，矩阵沿 Y 方向求和，折叠得到行向量 $p_X(x)$ 。行向量 $p_X(x)$ 进一步求和结果为标量 1，对应样本空间概率。反向来看，概率 1 沿 X 和 Y 展开，相当于“切片、切丝”。这个几何视角很重要，本章最后还要聊这个视角。

举个例子

如图 16 所示，当 $X=6$ 时，将整个一列的 $p_{X,Y}(6,y)$ 求和得到 $p_X(6)=0.2965$ 。请大家自己验算当 X 取其他值时，边缘概率 $p_X(x)$ 的具体值。

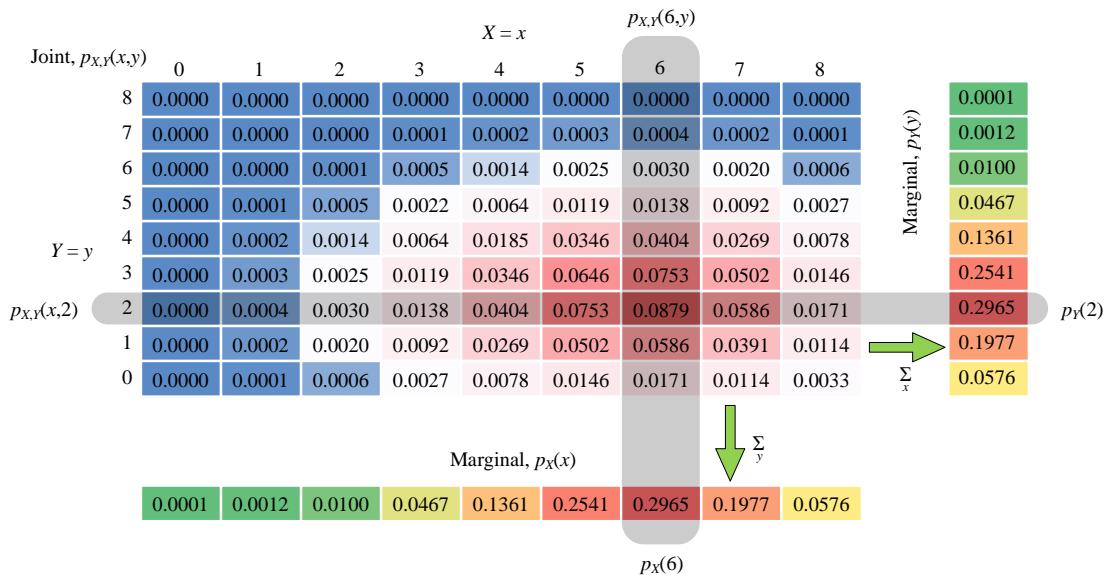


图 16. 利用联合概率计算边缘概率

边缘概率 $p_y(y)$

同理, $p_{x,y}(x, y)$ 对 x “偏求和”消去 x 得到 $p_y(y)$:

$$p_Y(y) = \sum_{\text{Marginal } x} p_{X,Y}(x, y)$$

(53)

图 16 所示, 当 $Y=2$ 时, 将整个一行的 $p_{x,y}(x, 2)$ 相加得到 $p_Y(2)=0.2965$ 。

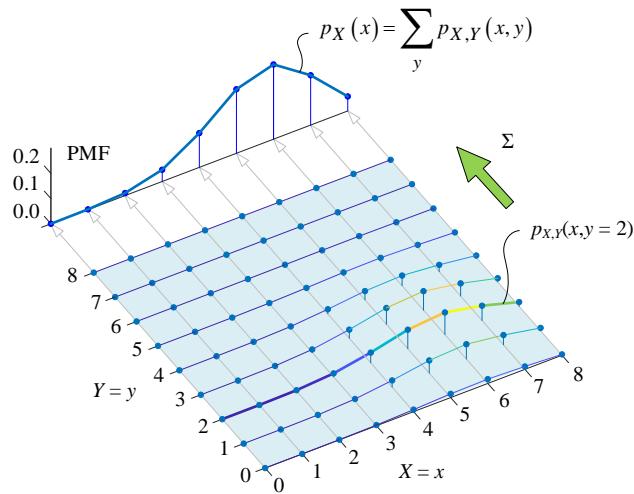
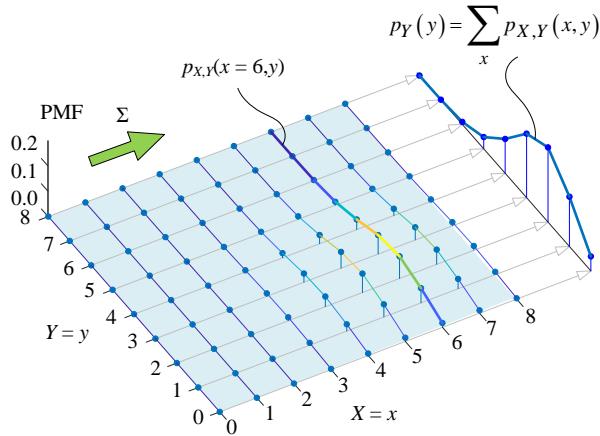
从函数角度来看, $p_Y(y)$ 也是个一元离散函数。

从矩阵运算角度来看, 矩阵 $p_{x,y}(x, y)$ 沿 X 方向求和, 折叠得到列向量 $p_Y(y)$ 。这相当于从二维降维到一维。

列向量 $p_Y(y)$ 进一步折叠结果同样为标量 1。

几何视角：叠加

显然, 边缘分布 $p_X(x)$ 和 $p_Y(y)$ 本身也是概率质量函数。从图像上来看, $p_X(x)$ 相当于 $p_{x,y}(x, y)$ 中 y 在取不同值时对应的火柴梗图叠加得到, 具体如图 17 所示。同理, 图 18 所示为边缘分布 $p_Y(y)$ 求解过程。

图 17. 边缘分布 $p_X(x)$ 求解过程图 18. 边缘分布 $p_Y(y)$ 求解过程

4.7 条件概率：引入贝叶斯定理

本节利用贝叶斯定理，介绍如何求解离散随机变量的条件概率质量函数。

联合概率 $p_{X,Y}(x,y) \rightarrow$ 条件概率 $p_{X|Y}(x|y)$

假设事件 $\{Y = y\}$ 已经发生，即 $p_Y(y) > 0$ 。在给定事件 $\{Y = y\}$ 条件下，事件 $\{X = x\}$ 发生的概率可以用条件概率质量函数 $p_{X|Y}(x|y)$ 表达。也就是说，对于 $p_{X|Y}(x|y)$ ， $\{Y = y\}$ 是新的样本空间。

利用贝叶斯定理，条件概率 $p_{X|Y}(x|y)$ 可以用联合概率 $p_{X,Y}(x,y)$ 除以边缘概率 $p_Y(y)$ 得到：

$$p_{X|Y}(x|y) = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}}$$
(54)

从函数角度来看， $p_{X|Y}(x|y)$ 本质上也是个二元函数。首先， $p_{X|Y}(x|y)$ 显然随着 $X=x$ 变化。虽然 $Y=y$ 为条件，但是这个条件也可以变动。 $Y=y$ 变动就会导致概率质量函数 $p_{X|Y}(x|y)$ 变化。

从矩阵运算角度来看， $p_{X,Y}(x,y)$ 相当于矩阵， $p_Y(y)$ 相当于列向量。两者相除用到《矩阵力量》第 4 章讲的**广播原则** (broadcasting)。得到的条件概率 $p_{X|Y}(x|y)$ 也是个矩阵，形状和 $p_{X,Y}(x,y)$ 一致。

$p_{X|Y}(x|y)$ 对 x 求和等于 1：

$$\sum_x p_{X|Y}(x|y) = 1$$
(55)

也就是说， $p_{X|Y}(x|y)$ 矩阵的每一行求和结果为 1。也就是说，每一行代表一个不同的“样本空间”。

注意，(55) 的结果实际上是一维数组， $\sum_x()$ 完成 X 方向压缩，但是 Y 这个维度没有被压缩。

换个视角来看，条件概率的“条件”就是“新的样本空间”，这个新的样本空间对应概率为 1。

举个例子

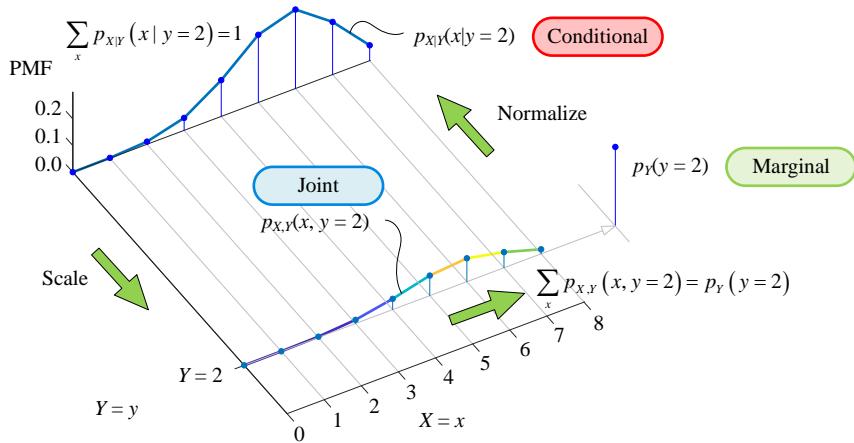
如图 19 所示， $Y=2$ 时，边缘概率 $p_Y(Y=2)$ 可以通过求和得到：

$$p_Y(2) = \sum_x p_{X,Y}(x, 2) \quad (56)$$

$p_Y(2)$ 为一定值。给定 $Y=2$ 作为条件时，条件概率 $p_{X|Y}(x|2)$ 通过下式得到：

$$p_{X|Y}(x|2) = \frac{\overbrace{p_{X,Y}(x,2)}^{\text{Joint}}}{\underbrace{p_Y(2)}_{\text{Marginal}}}$$
(57)

观察图 19，发现 $p_{X,Y}(x,2)$ 到 $p_{X|Y}(x|2)$ 相当于曲线缩放过程。

图 19. 求解条件概率 $p_{X|Y}(x|y)$ 的过程

进一步，条件概率 $p_{X|Y}(x|2)$ 对 x 求和得到 1：

$$\sum_x p_{X|Y}(x|2) = \frac{\sum_x p_{X,Y}(x,2)}{p_Y(2)} = \frac{p_Y(2)}{p_Y(2)} = 1 \quad (58)$$

$p_{X,Y}(x,2)$ 到 $p_{X|Y}(x|2)$ 是一个归一化 (normalization) 过程。也就是说，上式分母中的 $p_Y(2)$ 是一个归一化系数。这样，满足了归一化条件， $p_{X|Y}(x|2)$ 就“摇身一变”成了概率质量函数。

引入贝叶斯定理，边缘概率 $p_X(x)$ 相当于条件概率的加权平均：

$$p_X(x) = \underbrace{\sum_y p_{X,Y}(x,y)}_{\text{Marginal}} = \sum_y \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{p_Y(y)}_{\text{Marginal}} \quad (59)$$

条件概率 $p_{X|Y}(x|y) \rightarrow$ 联合概率 $p_{X,Y}(x,y)$

相反，条件概率 $p_{X|Y}(x|y)$ 到联合概率 $p_{X,Y}(x,y)$ 相当于，以边缘概率 $p_Y(y)$ 作为系数缩放 $p_{X|Y}(x|y)$ 的过程：

$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{p_Y(y)}_{\text{Marginal}} \quad (60)$$

条件概率 $p_{Y|X}(y|x)$

同理，给定事件 $\{X=x\}$ 条件下，当 $p_X(x) > 0$ ，事件 $\{Y=y\}$ 发生的概率可以用条件概率质量函数 $p_{Y|X}(y|x)$ 表达：

$$p_{Y|X}(y|x) = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_X(x)}_{\text{Marginal}}}$$
(61)

图 20 展示求解条件概率 $p_{Y|X}(y|x)$ 过程。同样，从函数角度来看， $p_{Y|X}(y|x)$ 也是个二元函数。从矩阵运算角度，上式也用到了广播原则，结果 $p_{Y|X}(y|x)$ 同样是个矩阵。

$p_{Y|X}(y|x)$ 对 y 求和等于 1：

$$\sum_y p_{Y|X}(y|x) = 1$$
(62)

也请大家从降维压缩角度理解上式。

(61) 也可以用来反求联合概率 $p_{Y,X}(y,x)$ ：

$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} \cdot \underbrace{p_X(x)}_{\text{Marginal}}$$
(63)

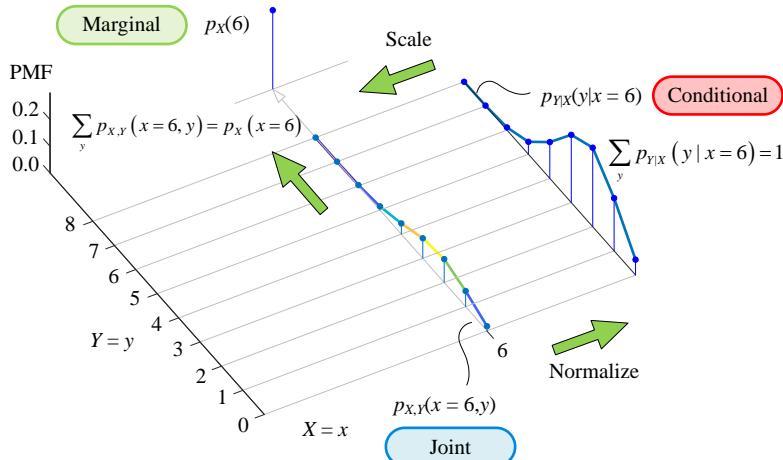


图 20. 求解条件概率 $p_{Y|X}(y|x)$ 的过程

同理，边缘概率 $p_Y(y)$ 也是条件概率 $p_{Y|X}(y|x)$ 的加权平均：

$$\underbrace{p_Y(y)}_{\text{Marginal}} = \sum_x p_{X,Y}(x,y) = \sum_y \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} \underbrace{p_X(x)}_{\text{Marginal}}$$
(64)

上式也是一个“偏求和”过程。

4.8 独立性：条件概率等于边缘概率

独立

如果两个离散变量 X 和 Y 独立，条件概率 $p_{X|Y}(x|y)$ 等于边缘概率 $p_X(x)$ ，下式成立：

$$\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} = \underbrace{p_X(x)}_{\text{Marginal}} \quad (65)$$

如图 21 所示， X 和 Y 独立，不管 y 取任何值 ($0 \sim 8$)， $p_X(x)$ 的形状和 $p_{X|Y}(x|y)$ 相同。

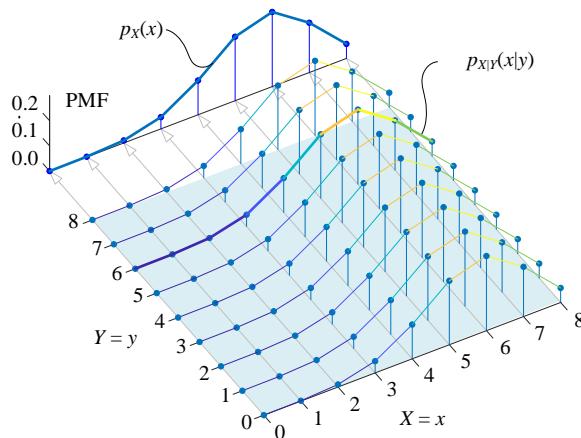


图 21. X 和 Y 独立，条件概率 $p_{X|Y}(x|y)$ 等于边缘概率 $p_X(x)$

(65) 等价于下式：

$$\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = \underbrace{p_Y(y)}_{\text{Marginal}} \quad (66)$$

同理，如图 22 所示， X 和 Y 独立时， $p_Y(y)$ 的形状和 $p_{Y|X}(y|x)$ 相同。这恰恰说明， X 的取值和 Y 无关，也就是为什么条件概率 $p_{Y|X}(y|x)$ 的形状不受 $X=x$ 影响，都和 $p_Y(y)$ 相同。

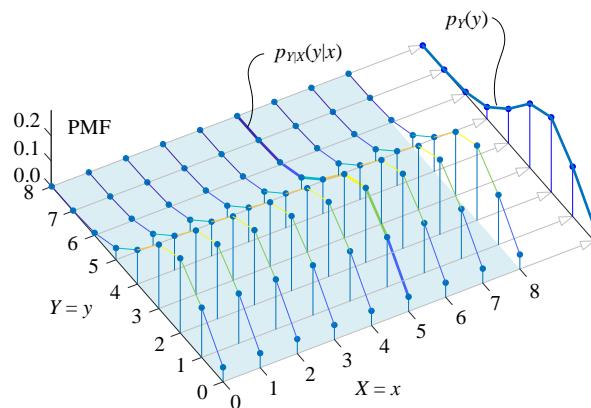


图 22. X 和 Y 独立，条件概率 $p_{Y|X}(y|x)$ 等于边缘概率 $p_Y(y)$

独立：计算联合概率 $p_{X,Y}(x,y)$

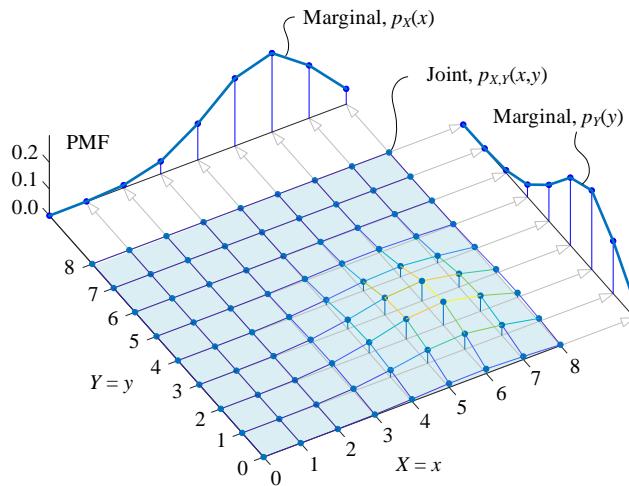
另外一个角度，如果离散随机变量 X 和 Y 独立，联合概率 $p_{X,Y}(x,y)$ 等于 $p_Y(y)$ 和 $p_X(x)$ 两个边缘概率质量函数 PMF 乘积：



$$\underbrace{p_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{p_Y(y)}_{\text{Marginal}} \cdot \underbrace{p_X(x)}_{\text{Marginal}}$$

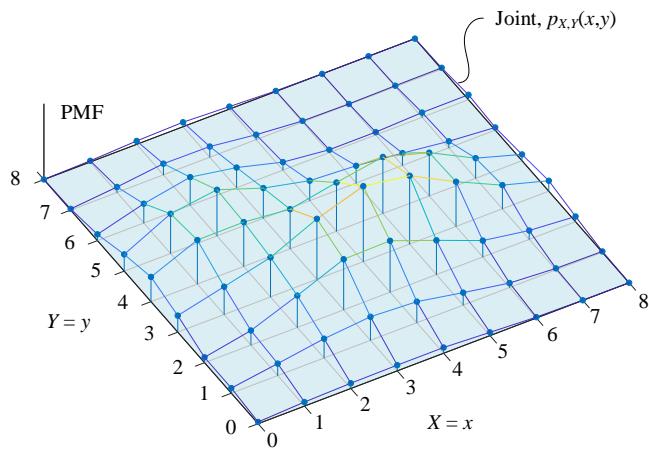
(67)

从向量角度来看，把 $p_Y(y)$ 和 $p_X(x)$ 看成是两个向量，上式相当于 $p_Y(y)$ 和 $p_X(x)$ 的张量积。

图 23. 联合概率 $p_{X,Y}(x,y)$ 等于 $p_Y(y)$ 和 $p_X(x)$ 两个边缘概率乘积

不独立

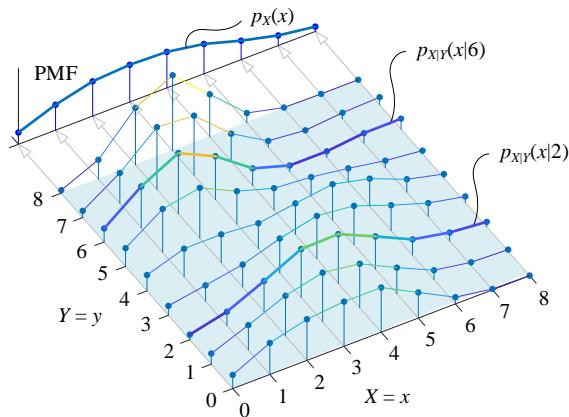
我们再来看一下，在离散随机变量 X 和 Y 不独立的情况下， $p_{Y|X}(y|x)$ 和 $p_Y(y)$ 图像可能存在的某种关系。图 24 给出另一个联合概率 $p_{X,Y}(x,y)$ 的图像。

图 24. 离散随机变量 X 和 Y 不独立情况下，联合概率 $p_{X,Y}(x,y)$

前文已经介绍，如果 X 和 Y 不独立，如果 $p_Y(y) > 0$ ，条件概率 $p_{X|Y}(x|y)$ 公式如下：

$$\underbrace{p_{X|Y}(x|y)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\sum_x p_{X,Y}(x,y)} \quad (68)$$

如图 25 所示，当 X 和 Y 不独立，条件概率 $p_{X|Y}(x|y)$ 不同于边缘概率 $p_X(x)$ 。

图 25. X 和 Y 不独立，条件概率 $p_{X|Y}(x|y)$ 不同于边缘概率 $p_X(x)$

如果 $p_X(x) > 0$ ，条件概率 $p_{Y|X}(y|x)$ 需要利用贝叶斯定理计算：

$$\underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_X(x)}_{\text{Marginal}}} = \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\sum_y p_{X,Y}(x,y)} \quad (69)$$

如图 26 所示， X 和 Y 不独立，条件概率 $p_{Y|X}(y|x)$ 不同于边缘概率 $p_Y(y)$ 。

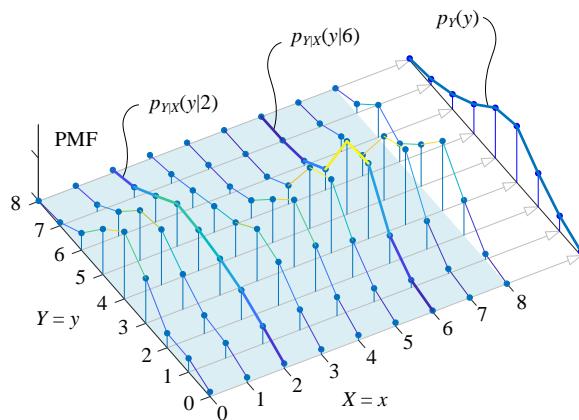


图 26. X 和 Y 不独立，条件概率 $p_{Y|X}(y|x)$ 不同于边缘概率 $p_Y(y)$

4.9 以鸢尾花数据为例：不考虑分类标签

本章下两节用鸢尾花数据集花萼长度 (X_1)、花萼宽度 (X_2)、分类标签 (Y) 样本数据为例，讲解离散随机变量主要知识点。

对于鸢尾花数据集，分类标签 (Y) 本身就是离散随机变量，因为 Y 的取值只有三个，对应鸢尾花三个类别——versicolor、setosa、virginica。

而花萼长度 (X_1)、花萼宽度 (X_2) 两者取值都是连续数值，大家可能好奇， X_1 和 X_2 怎么可能变成离散随机变量？

两把直尺

这里只需要做一个很小的调整，给定鸢尾花花萼长度或宽度 d ，然后进行 $\text{round}(2 \times d)/2$ 运算。比如，鸢尾花花萼长度为 5.3，进行上述计算变成 5.5。

这就好比，测量鸢尾花获得原始数据时，用的是图 27 (a) 所示直尺。而我们在测量花萼长度、花萼宽度时，用的是如图 27 (b) 所示的直尺。直尺精度为 0.5 cm。而测量结果仅保留一位有效小数，这一位小数的数值可能是 0 或 5。

实际上鸢尾花四个特征的原始数据本身也是“离散的”，因为原始数据仅仅保留一位有效小数位。只不过我们把数据看成是连续数据而已。从这个角度来看，在数据科学领域，电子数据离散、连续与否是相对的。

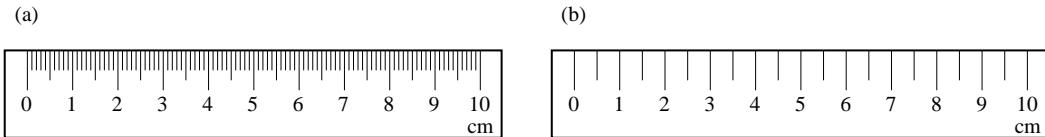


图 27. 两把直尺

“离散”的花萼长度、花萼宽度数据

图 28 所示为经过 $\text{round}(2 \times d)/2$ 运算得到的“离散”的花萼长度、花萼宽度数据散点图。

花萼长度 (X_1) 取值有 8 个，分别是 4.5、5.0、5.5、6.0、6.5、7.0、7.5、8.0。也就是说 X_1 的样本空间为 $\{4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5, 8.0\}$ 。

花萼宽度 (X_2) 取值有 6 个，分别是 2.0、2.5、3.0、3.5、4.0、4.5。 X_2 的样本空间为 $\{2.0, 2.5, 3.0, 3.5, 4.0, 4.5\}$ 。

下一步，我们统计每个散点对应的频数，即散点图中网格线交点处样本数量。

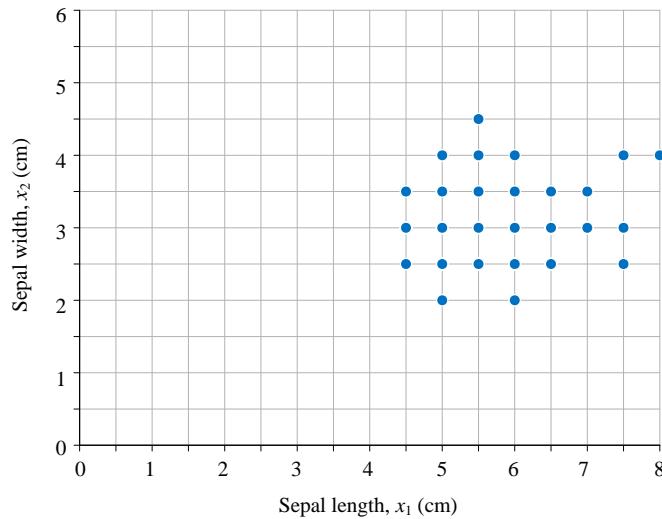


图 28. “离散”的鸢尾花花萼长度、花萼宽度散点图

频数 → 联合概率质量函数 $p_{X1,X2}(x_1, x_2)$

基于图 28 所示数据，我们可以得到图 29 所示频数和概率热图。为了区分频数和概率热图，两类热图采用不同色谱。

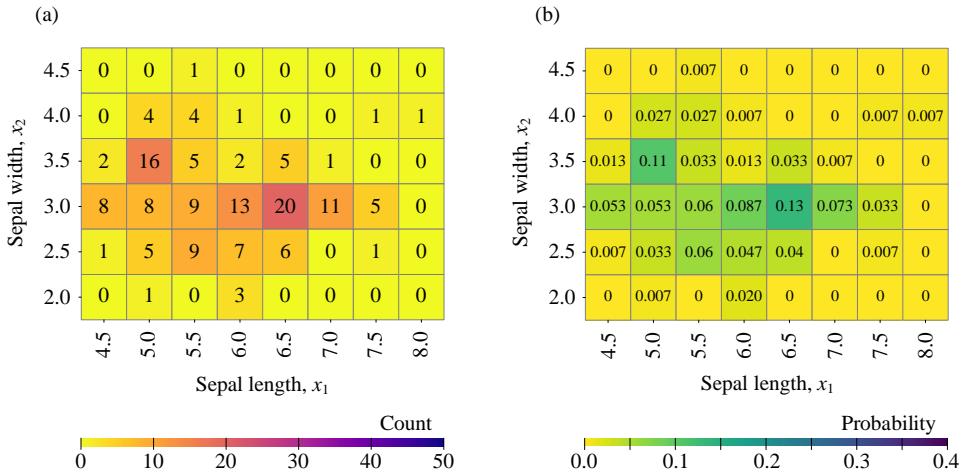


图 29. 频数和概率热图，全部样本点，不考虑分类

图 29 (a) 中频数之和为 150，即鸢尾花样本总数。从频数到概率的计算很简单，比如频数为 3，样本总数为 150，两者比值对应概率 $0.02 = 3/150$ 。

翻译成“概率语言”就是，根据既有样本数据，花萼长度 (X_1) 为 6.0、花萼宽度 (X_2) 为 2.0 对应的联合概率为 0.02：

$$p_{X_1, X_2}(6.0, 2.0) = 0.02 \quad (70)$$

采用穷举法，图 29 (b) 热图中所有取值之和为 1，即：

$$\sum_{x_1} \sum_{x_2} p_{X_1, X_2}(x_1, x_2) = 1 \quad (71)$$

用样本数来计算的话，上式相当于 $150/150 = 1$ 。也就是说，图 29 (b) 是对概率为 1 的某种特定的分割。

花萼长度边缘概率 $p_{X_1}(x_1)$ ：偏求和

图 30 所示为求解花萼长度边缘概率的过程。

举个例子，当花萼长度 (X_1) 取值为 7.0 时，对应的边缘概率 $p_{X_1}(7.0)$ 可以通过如下“偏求和”得到：

$$p_{X_1}(7.0) = \sum_{x_2} p_{X_1, X_2}(7.0, x_2) = \frac{0}{X_2=2.0} + \frac{0}{X_2=2.5} + \frac{0.073}{X_2=3.0} + \frac{0.007}{X_2=3.5} + \frac{0}{X_2=4.0} + \frac{0}{X_2=4.5} = 0.08 \quad (72)$$

上式相当于，固定花萼长度 (X_1) 为 7.0，然后穷举花萼宽度 (X_2) 所有概率值，然后求和（压缩、折叠）。

从频数角度来看，上式相当于：

$$X_2=2.0 \quad X_2=2.5 \quad X_2=3.0 \quad X_2=3.5 \quad X_2=4.0 \quad X_2=4.5 \\ p_{X_1}(7.0) = \frac{0 + 0 + 11 + 1 + 0 + 0}{150} = \frac{12}{150} = 0.08 \quad (73)$$

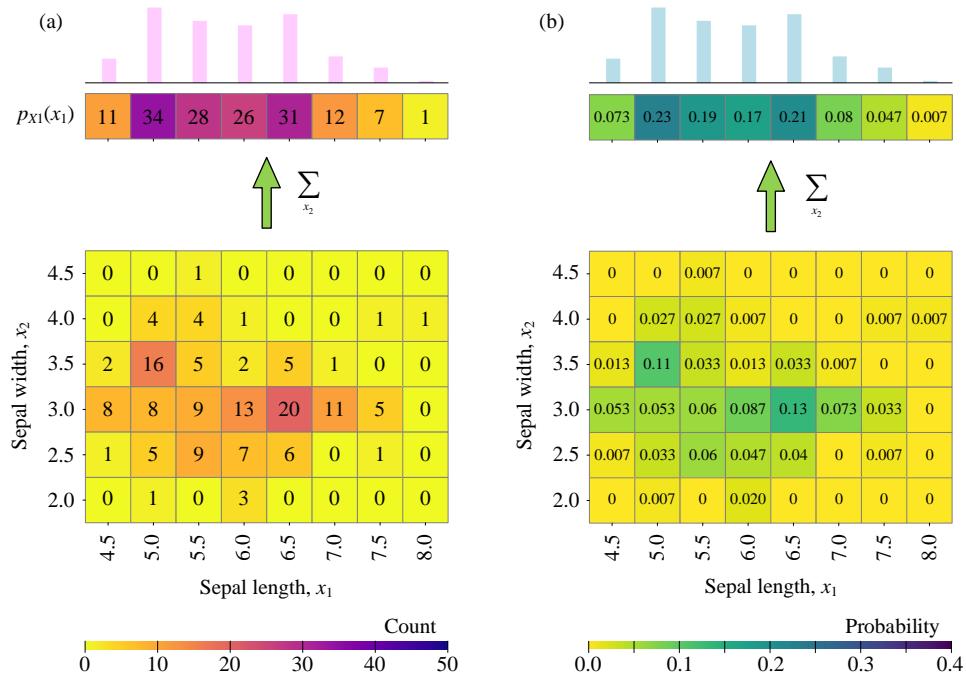


图 30. 花萼长度的边缘频数和概率热图，不考虑分类

花萼宽度边缘概率 $p_{X_2}(x_2)$: 偏求和

图 31 所示为求解花萼宽度边缘概率的过程。

举个例子，当花萼宽度 (X_2) 取值为 2.0 时，对应的边缘概率 $p_{X_2}(2.0)$ 可以通过如下偏求和得到：

$$p_{X_2}(2.0) = \sum_{x_1} p_{X_1, X_2}(x_1, 2.0) = \underset{X_1=4.5}{0} + \underset{X_1=5.0}{0.007} + \underset{X_1=5.5}{0} + \underset{X_1=6.0}{0.02} + \underset{X_1=6.5}{0} + \underset{X_1=7.0}{0} + \underset{X_1=7.5}{0} + \underset{X_1=8.0}{0} = 0.027 \quad (74)$$

上式相当于，固定花萼宽度 (X_2) 为 2.0，然后穷举花萼长度 (X_1) 所有概率值，然后求和。

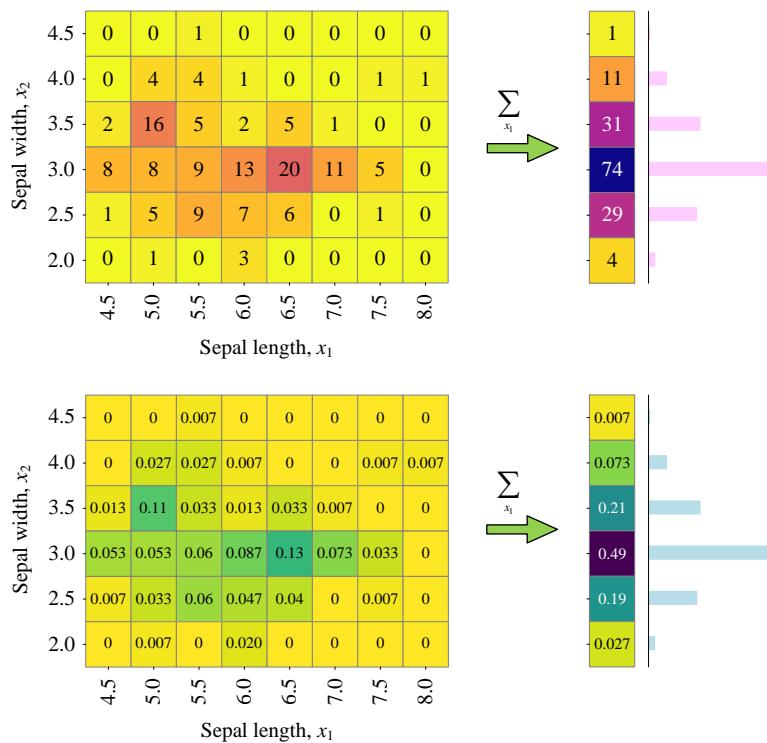


图 31. 花萼宽度的边缘频数和概率热图，不考虑分类

期望值、方差

花萼长度 X_1 的期望值：

$$\begin{aligned}
 E(X_1) &= \sum_{x_1} x_1 \cdot p_{X_1}(x_1) \\
 &= 4.5 \times 0.073 + 5.0 \times 0.23 + 5.5 \times 0.19 + 6.0 \times 0.17 + \\
 &\quad 6.5 \times 0.21 + 7.0 \times 0.08 + 7.5 \times 0.047 + 8.0 \times 0.007 \\
 &= 5.836 \text{ cm}
 \end{aligned} \tag{75}$$

请大家自行写出上式对应的矩阵运算式，并画出矩阵乘法运算示意图。

然后，计算花萼长度 X_1 平方的期望值：

$$\begin{aligned}
 E(X_1^2) &= \sum_{x_1} x_1^2 \cdot p_{X_1}(x_1) \\
 &= 4.5^2 \times 0.073 + 5.0^2 \times 0.23 + 5.5^2 \times 0.19 + 6.0^2 \times 0.17 + \\
 &\quad 6.5^2 \times 0.21 + 7.0^2 \times 0.08 + 7.5^2 \times 0.047 + 8.0^2 \times 0.007 \\
 &= 34.741 \text{ cm}^2
 \end{aligned} \tag{76}$$

由此可以求得花萼长度 X_1 的方差：

$$\text{var}(X_1) = \underbrace{\mathbb{E}(X_1^2)}_{\text{Expectation of } X^2} - \underbrace{\mathbb{E}(X_1)^2}_{\text{Square of } \mathbb{E}(X_1)} = 0.6749 \quad (77)$$

结果的单位为平方厘米，cm²。

⚠ 注意：上式把数据当做总体的样本数据看待。

(77) 的平方根便是 X_1 的标准差：

$$\sigma_{x_1} = \sqrt{\text{var}(X_1)} = 0.821 \text{ cm} \quad (78)$$

请大家自行计算：花萼宽度 X_2 的期望值、 X_2 平方期望值。由此，可以求得花萼宽度 X_2 的方差，然后计算 X_2 的标准差。

独立

前文提过，如果假设 X_1 和 X_2 独立，联合概率可通过下式计算得到：

$$p_{X_1, X_2}(x_1, x_2) = p_{X_1}(x_1) \cdot p_{X_2}(x_2) \quad (79)$$

图 32 所示为，假设 X_1 和 X_2 独立，联合概率的热图。

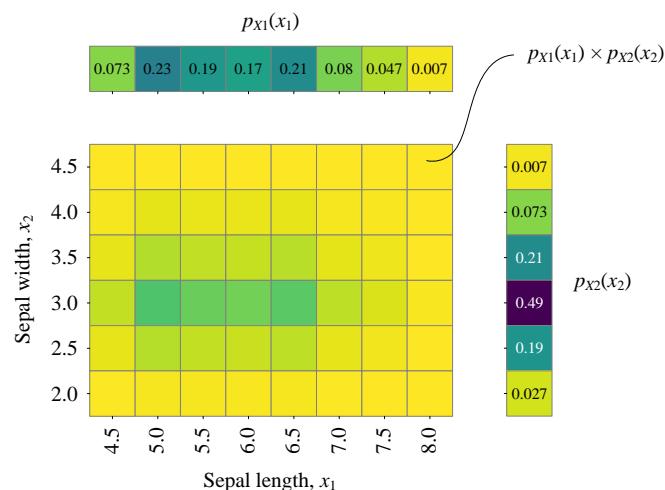


图 32. 联合概率，假设独立

这实际上就是《矩阵力量》介绍的向量张量积，也相当于如图 33 所示的矩阵乘法。

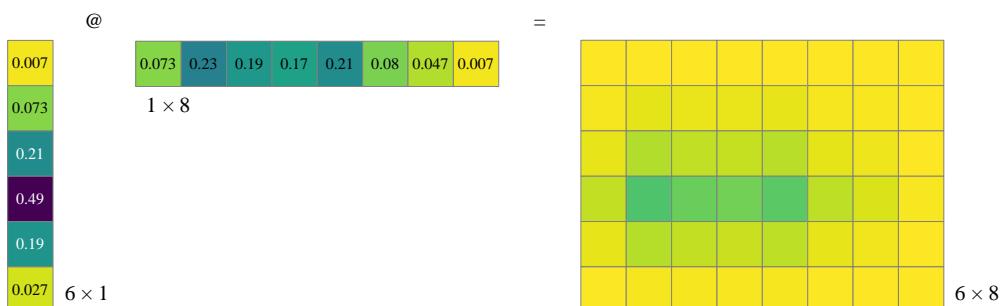


图 33. X_1 和 X_2 条件独立，矩阵乘法

图 32 中矩阵所有元素之和也是 1。追根溯源，这体现的是乘法的分配律：

$$\underbrace{\sum_{x_1} p_{X_1}(x_1)}_{=1} \cdot \underbrace{\sum_{x_2} p_{X_2}(x_2)}_{=1} = 1 \quad (80)$$

为了配合热图形式，用如下方式展开上式：

$$\underbrace{\{p_{X_2}(4.5) + p_{X_2}(4.0) + \dots + p_{X_2}(2.0)\}}_{=1} \cdot \underbrace{\{p_{X_1}(4.5) + p_{X_1}(5.0) + \dots + p_{X_1}(8.0)\}}_{=1} = 1 \quad (81)$$

展开的每一个元素对应热图矩阵的每个元素：

$$\begin{aligned} & p_{X_2}(4.5) \cdot p_{X_1}(4.5) + p_{X_2}(4.5) \cdot p_{X_1}(5.0) + \dots + p_{X_2}(4.5) \cdot p_{X_1}(8.0) + \\ & p_{X_2}(4.0) \cdot p_{X_1}(4.5) + p_{X_2}(4.0) \cdot p_{X_1}(5.0) + \dots + p_{X_2}(4.0) \cdot p_{X_1}(8.0) + \\ & \dots + \\ & p_{X_2}(2.0) \cdot p_{X_1}(4.5) + p_{X_2}(2.0) \cdot p_{X_1}(5.0) + \dots + p_{X_2}(2.0) \cdot p_{X_1}(8.0) = 1 \end{aligned} \quad (82)$$

比较图 32 和图 29 (b)，我们发现假设 X_1 和 X_2 独立得到的联合概率和真实值偏差很大。也就是说，(79) 这种假设随机变量独立然后计算联合概率的方法很多时候并不准确，需要谨慎。

给定花萼长度，花萼宽度的条件概率 $p_{X_2|X_1}(x_2 | x_1)$

如图 34 所示，给定花萼长度 $X_1 = 5.0$ 作为条件，这相当于在整个样本空间中，单独划出一个区域（浅蓝色）。这个区域将是“条件概率样本空间”，对应图 34 中的浅蓝色背景区域。计算 $X_1 = 5.0$ 条件概率时，将浅蓝色区域的概率值设为 1。

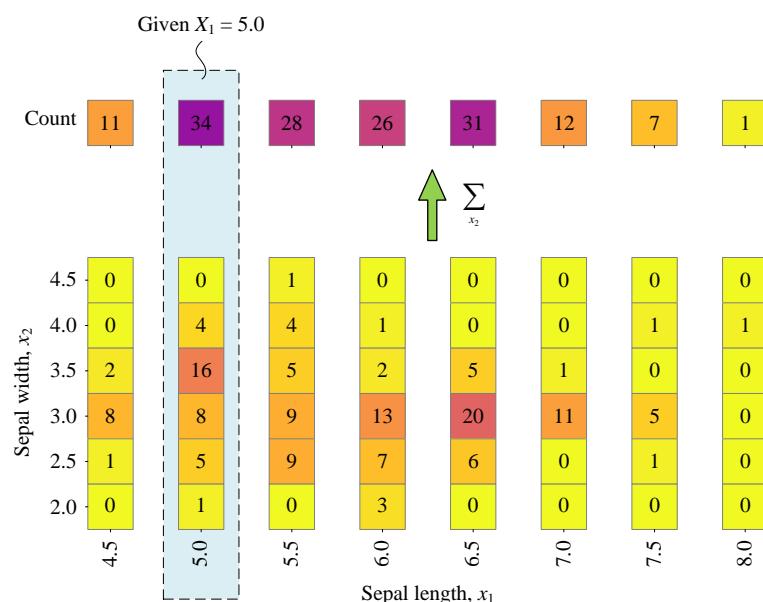


图 34. 频数视角，给定花萼长度，如何计算花萼宽度的条件概率

采用穷举法，这个区域中的条件概率有如下几个：

$$\begin{aligned}
 p_{X_2|X_1}(x_2 = 4.5 | x_1 = 5.0) &= \frac{0}{34} = 0 \\
 p_{X_2|X_1}(x_2 = 4.0 | x_1 = 5.0) &= \frac{4}{34} \approx 0.12 \\
 p_{X_2|X_1}(x_2 = 3.5 | x_1 = 5.0) &= \frac{16}{34} \approx 0.47 \\
 p_{X_2|X_1}(x_2 = 3.0 | x_1 = 5.0) &= \frac{8}{34} \approx 0.24 \\
 p_{X_2|X_1}(x_2 = 2.5 | x_1 = 5.0) &= \frac{5}{34} \approx 0.15 \\
 p_{X_2|X_1}(x_2 = 2.0 | x_1 = 5.0) &= \frac{1}{34} \approx 0.029
 \end{aligned} \tag{83}$$

换个方法来求。如图 35 所示，利用贝叶斯定理，(83) 中条件概率可以通过下式计算：

$$\begin{aligned}
 p_{X_2|X_1}(x_2 = 4.5 | x_1 = 5.0) &= \frac{p_{X_1, X_2}(x_1 = 5.0, x_2 = 4.5)}{p_{X_1}(x_1 = 5.0)} \approx \frac{0}{0.23} = 0 \\
 p_{X_2|X_1}(x_2 = 4.0 | x_1 = 5.0) &= \frac{p_{X_1, X_2}(x_1 = 5.0, x_2 = 4.0)}{p_{X_1}(x_1 = 5.0)} \approx \frac{0.027}{0.23} \approx 0.12 \\
 p_{X_2|X_1}(x_2 = 3.5 | x_1 = 5.0) &= \frac{p_{X_1, X_2}(x_1 = 5.0, x_2 = 3.5)}{p_{X_1}(x_1 = 5.0)} \approx \frac{0.11}{0.23} \approx 0.47 \\
 p_{X_2|X_1}(x_2 = 3.0 | x_1 = 5.0) &= \frac{p_{X_1, X_2}(x_1 = 5.0, x_2 = 3.0)}{p_{X_1}(x_1 = 5.0)} \approx \frac{0.053}{0.23} \approx 0.24 \\
 p_{X_2|X_1}(x_2 = 2.5 | x_1 = 5.0) &= \frac{p_{X_1, X_2}(x_1 = 5.0, x_2 = 2.5)}{p_{X_1}(x_1 = 5.0)} \approx \frac{0.033}{0.23} \approx 0.15 \\
 p_{X_2|X_1}(x_2 = 2.0 | x_1 = 5.0) &= \frac{p_{X_1, X_2}(x_1 = 5.0, x_2 = 2.0)}{p_{X_1}(x_1 = 5.0)} \approx \frac{0.007}{0.23} \approx 0.029
 \end{aligned} \tag{84}$$

其中，

$$\begin{aligned}
 p_{X_1}(x_1 = 5.0) &= p_{X_1, X_2}(x_1 = 5.0, x_2 = 4.5) + p_{X_1, X_2}(x_1 = 5.0, x_2 = 4.0) + \\
 &\quad p_{X_1, X_2}(x_1 = 5.0, x_2 = 3.5) + p_{X_1, X_2}(x_1 = 5.0, x_2 = 3.0) + \\
 &\quad p_{X_1, X_2}(x_1 = 5.0, x_2 = 2.5) + p_{X_1, X_2}(x_1 = 5.0, x_2 = 2.0) \\
 &\approx 0 + 0.027 + 0.11 + 0.053 + 0.033 + 0.007 \approx 0.23
 \end{aligned} \tag{85}$$

比较 (83) 和 (84)，发现结果相同。

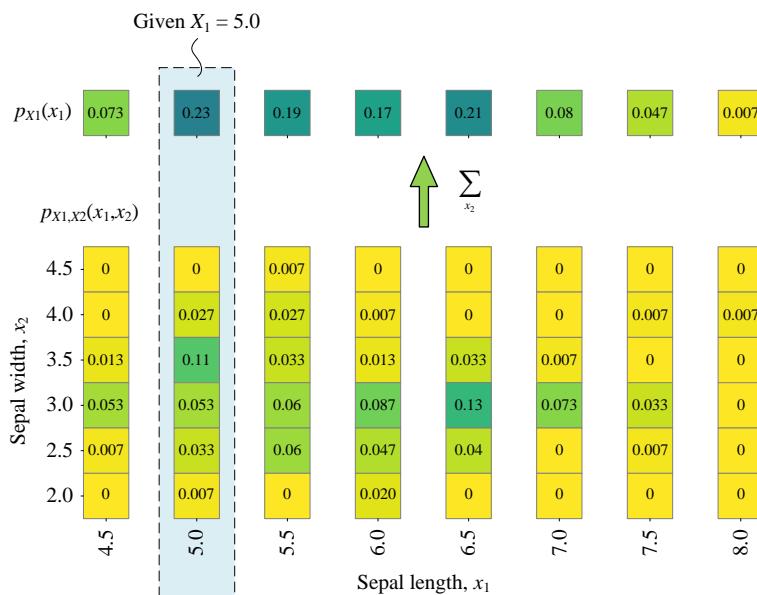
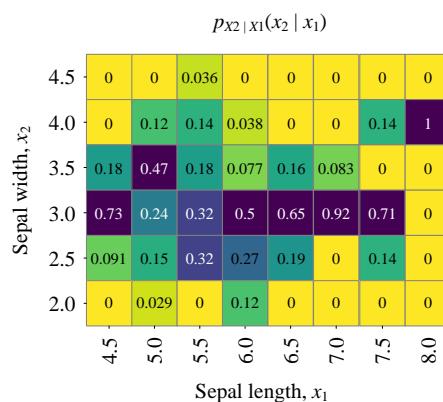


图 35. 概率视角，给定花萼长度，如何计算花萼宽度的条件概率

本章前文提过，从函数角度来看， $p_{X2|X1}(x_2|x_1)$ 本质上也是个二元离散函数，具体如图 36 所示。

图 36. 给定花萼长度，花萼宽度的条件概率 $p_{X2|X1}(x_2|x_1)$

如图 37 所示，每一列条件概率求和为 1：

$$\sum_{x_2} p_{X2|X1}(x_2|x_1) = 1 \quad (86)$$

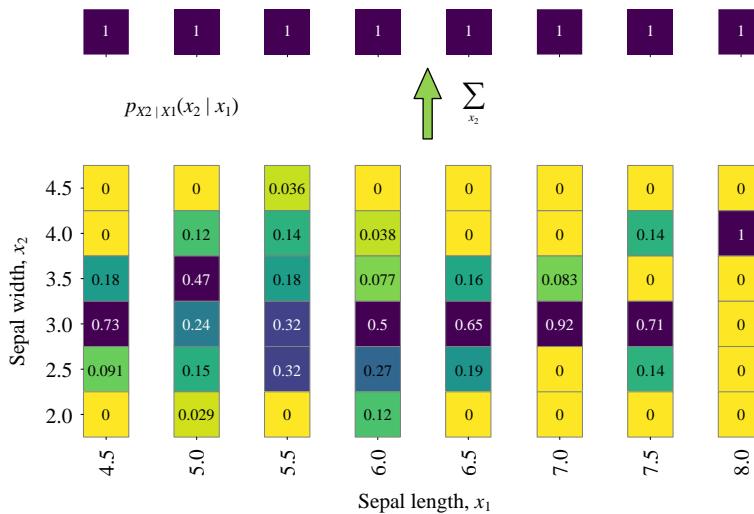


图 37. 给定花萼长度，花萼宽度的条件概率，每一列条件概率求和为 1

给定花萼宽度，花萼长度的条件概率 $p_{X1|X2}(x_1 | x_2)$

根据图 38 数据，请大家自行计算，给定花萼宽度为 3.0，每个条件概率 $p_{X1|X2}(x_1 | 3.0)$ 的具体值。

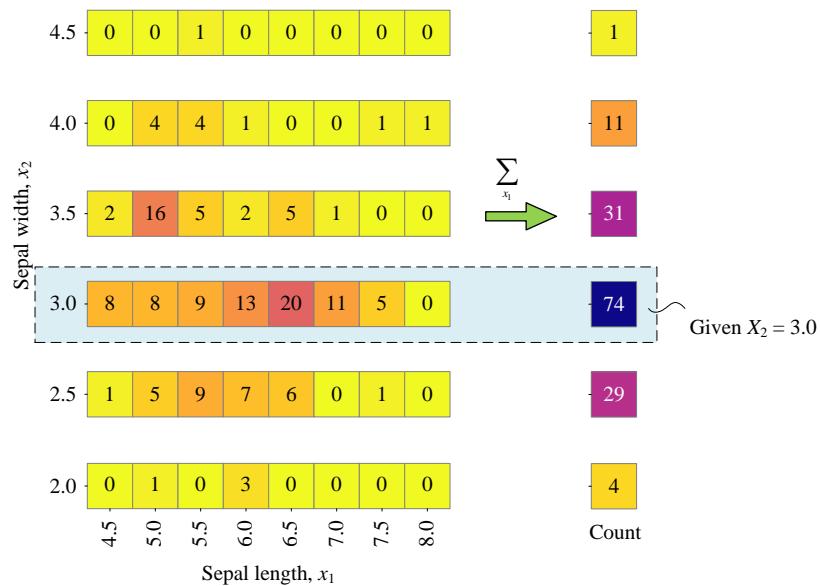


图 38. 频数视角，给定花萼宽度，如何计算花萼长度的条件概率

从函数角度来看， $p_{X1|X2}(x_1 | x_2)$ 也是个二元离散函数，具体如图 39 所示。

大家是否立刻想到，既然我们可以求得花萼长度的期望值，我们是否可以求得给定花萼宽度条件下的花萼长度的期望、方差？

答案是肯定的！

→ 本书第 8 章将专门介绍**条件期望** (conditional expectation)、**条件方差** (conditional variance)。

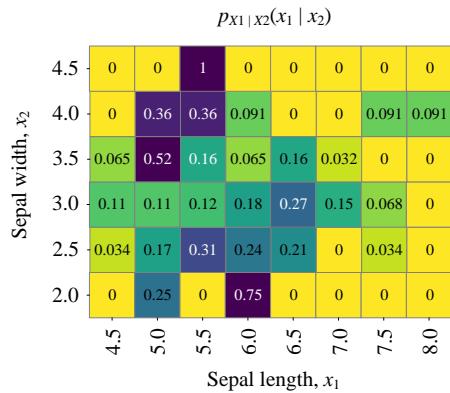


图 39. 给定花萼宽度，花萼长度的条件概率 $p_{X1|X2}(x_1 | x_2)$

如图 40 所示，每一行条件概率求和为 1：

$$\sum_{x_1} p_{X1|X2}(x_1 | x_2) = 1 \quad (87)$$

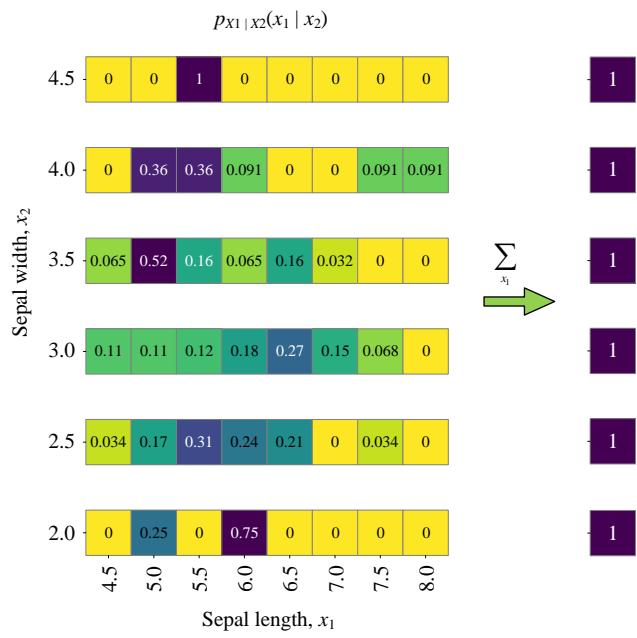


图 40. 给定花萼宽度，花萼长度的条件概率每一行条件概率求和为 1

4.10 以鸢尾花数据为例：考虑分类标签

本节讨论在考虑分类标签条件下，如何计算鸢尾花数据的条件概率。

给定分类标签 $Y = C_1$ (setosa)

图 41 (a) 所示为给定分类标签 $Y = C_1$ (setosa) 条件下，鸢尾花数据集中 50 个样本数据的频数热图。图 41 中频数除以 50 便得到图 41 (b) 所示条件概率 $p_{X_1, X_2 | Y}(x_1, x_2 | y = C_1)$ 热图。

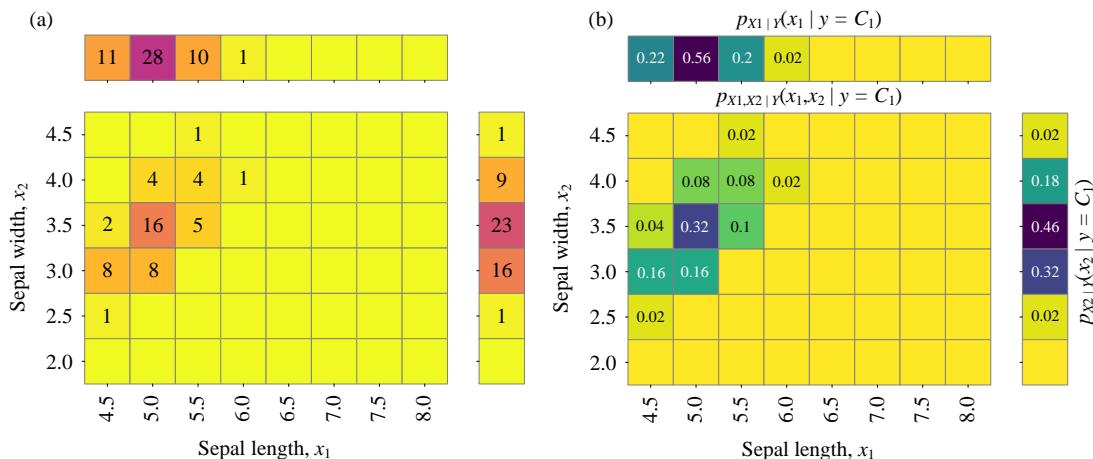


图 41. 频数和条件概率 $p_{X_1, X_2 | Y}(x_1, x_2 | y = C_1)$ 热图，给定分类标签 $Y = C_1$ (setosa)

此外，请大家根据频数热图，自行计算两个条件概率： $p_{X_1 | X_2, Y}(x_1 = 5.0 | x_2 = 3.0, y = C_1)$ 和 $p_{X_2 | X_1, Y}(x_2 = 3.0 | x_1 = 5.0, y = C_1)$ 。

给定分类标签 $Y = C_2$ (versicolor)

图 42 (a) 所示为给定分类标签 $Y = C_2$ (versicolor) 条件下，鸢尾花数据集中 50 个样本数据的频数热图。图 42 中频数除以 50 便得到图 42 (b) 所示条件概率 $p_{X_1, X_2 | Y}(x_1, x_2 | y = C_2)$ 热图。

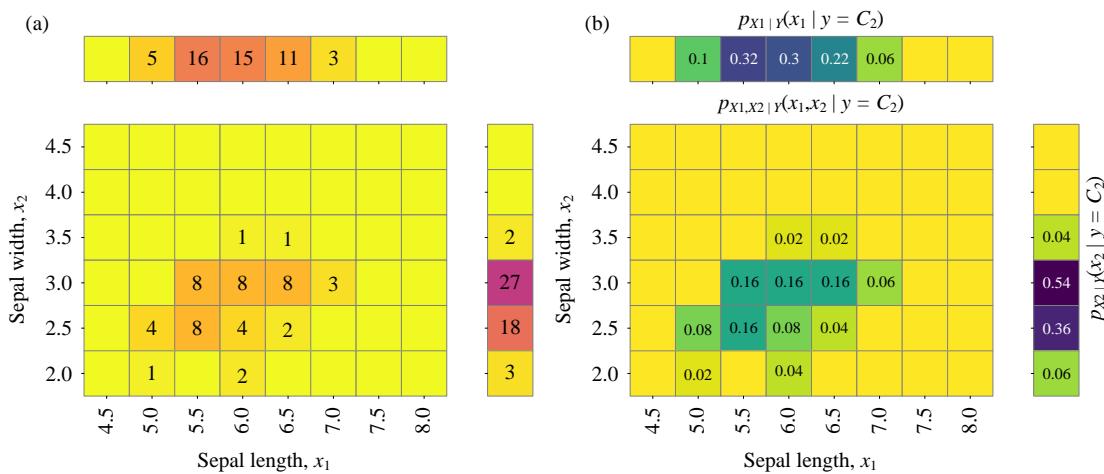


图 42. 频数和条件概率 $p_{X_1, X_2 | Y}(x_1, x_2 | y = C_2)$ 热图，给定分类标签 $Y = C_2$ (versicolor)

给定分类标签 $Y = C_3$ (virginica)

请大家自行分析图 43。

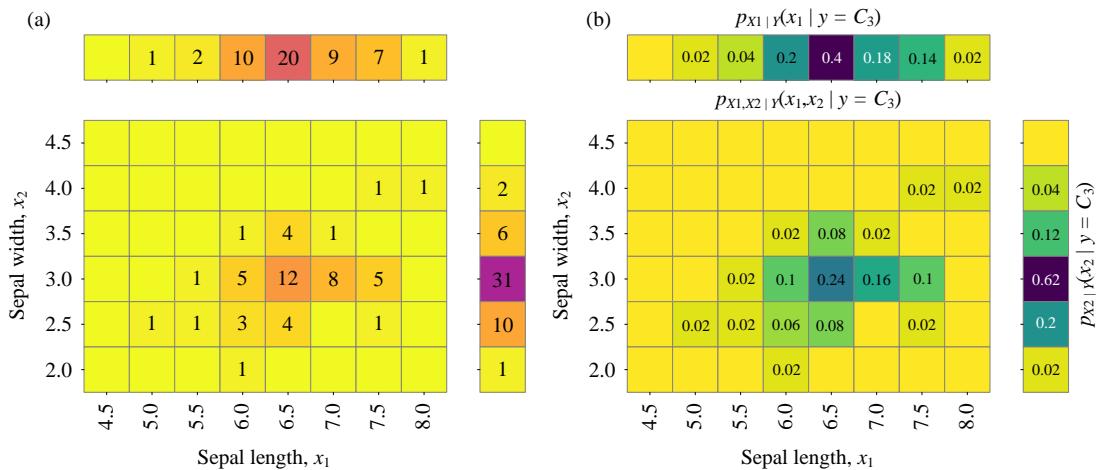


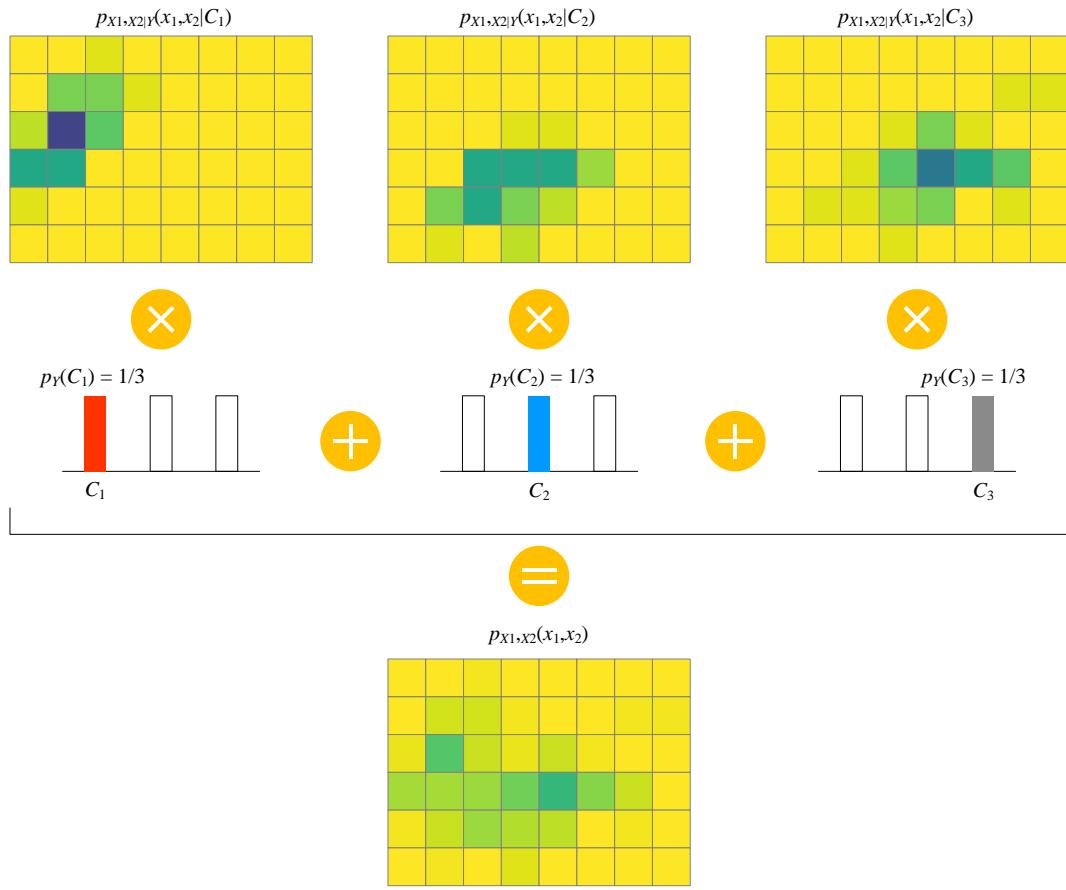
图 43. 频数和条件概率 $p_{X_1, X_2 | Y}(x_1, x_2 | y = C_3)$ 热图, 给定分类标签 $Y = C_3$ (virginica)

全概率

如图 44 所示, 利用全概率定理, 我们可以通过下式计算 $p_{X_1, X_2}(x_1, x_2)$:

$$\begin{aligned}
 p_{X_1, X_2}(x_1, x_2) &= \sum_y \underbrace{p_{X_1, X_2, Y}(x_1, x_2, y)}_{\text{Joint}} \\
 &= \sum_y \underbrace{p_{X_1, X_2|Y}(x_1, x_2 | y)}_{\text{Conditional}} \cdot \underbrace{p_Y(y)}_{\text{Marginal}} \\
 &= p_{X_1, X_2|Y}(x_1, x_2 | C_1) \cdot p_Y(C_1) + \\
 &\quad p_{X_1, X_2|Y}(x_1, x_2 | C_2) \cdot p_Y(C_2) + \\
 &\quad p_{X_1, X_2|Y}(x_1, x_2 | C_3) \cdot p_Y(C_3)
 \end{aligned} \tag{88}$$

从几何角度来看, 联合概率质量函数 $p_{X_1, X_2}(x_1, x_2, y)$ 相当于一个“立方体”。上式相当于, 将立方体在 Y 方向上压扁成 $p_{X_1, X_2}(x_1, x_2)$ 平面。本章最后将继续这一话题。

图 44. 利用全概率定理，计算 $p_{X1,X2}(x_1, x_2)$

条件独立

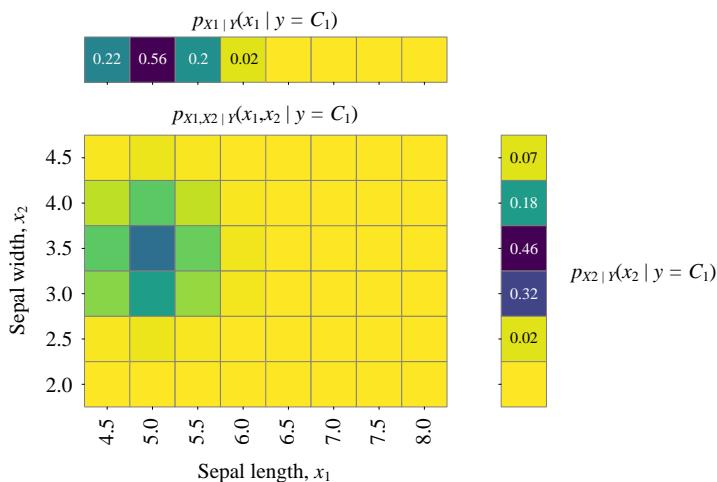
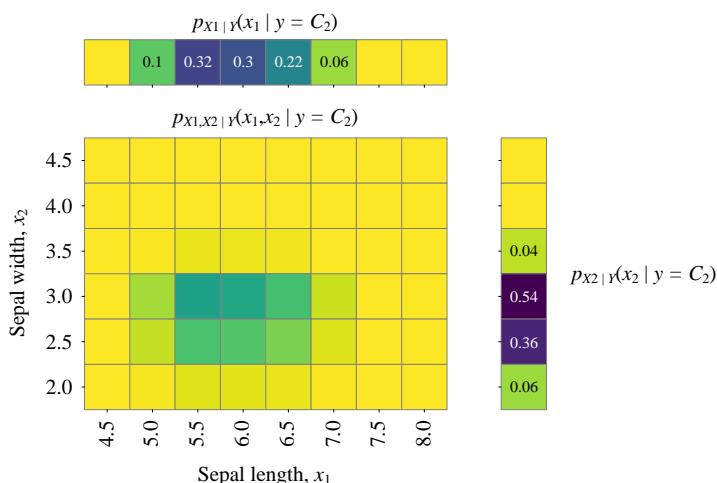
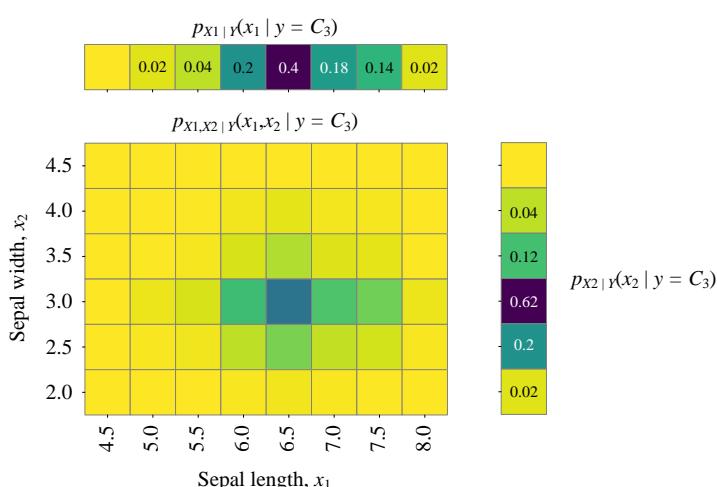
图 45 所示为给定 $Y = C_1$ 条件下，假设 X_1 和 X_2 条件独立，利用 $p_{X1|Y}(x_1 | y = C_1)$ 、 $p_{X2|Y}(x_2 | y = C_1)$ 估算 $p_{X1,X2|Y}(x_1, x_2 | y = C_1)$ ：

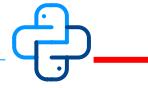
$$p_{X1,X2|Y}(x_1, x_2 | C_1) = p_{X1|Y}(x_1 | C_1) p_{X2|Y}(x_2 | C_1) \quad (89)$$

图 45 也相当于两个向量的张量积，请大家画出矩阵运算示意图。

请大家自行从矩阵乘法角度分析图 46、图 47。

将这些条件概率质量函数代入 (88)，我们也可以计算得到另外一个 $p_{X1,X2}(x_1, x_2)$ 。这实际上是估算 $p_{X1,X2}(x_1, x_2)$ 的一种方法。本书后续还会介绍这种方法及其应用。

图 45. 给定 $Y = C_1$, 假设 X_1 和 X_2 条件独立, 计算 $p_{X_1,X_2|Y}(x_1,x_2 | y = C_1)$ 图 46. 给定 $Y = C_2$, 假设 X_1 和 X_2 条件独立, 计算 $p_{X_1,X_2|Y}(x_1,x_2 | y = C_2)$ 图 47. 给定 $Y = C_3$, 假设 X_1 和 X_2 条件独立, 计算 $p_{X_1,X_2|Y}(x_1,x_2 | y = C_3)$



代码 Bk5_Ch04_02.py 绘制前两节大部分图像。

4.10 再谈概率 1：展开、折叠

偏求和：压扁

本章前文提到，几何上， $p_{X1,X2,X3}(x_1, x_2, x_3)$ 可以视作一个三维立方体。而偏求和是个降维过程，把立方体在不同维度上压扁。

如图 48 所示， $p_{X1,X2,X3}(x_1, x_2, x_3)$ 在 x_1 上偏求和，压扁得到 $p_{X2,X3}(x_2, x_3)$ ：

$$p_{X2,X3}(x_2, x_3) = \sum_{x_1} p_{X1,X2,X3}(x_1, x_2, x_3) \quad (90)$$

如图 48 所示， $p_{X2,X3}(x_2, x_3)$ 代表一个二维平面，相当于一个矩阵。

而 $p_{X2,X3}(x_2, x_3)$ 进一步沿着 x_2 折叠便得到边缘概率质量函数 $p_{X3}(x_3)$ ：

$$\begin{aligned} p_{X3}(x_3) &= \sum_{x_2} p_{X2,X3}(x_2, x_3) \\ &= \sum_{x_2} \sum_{x_1} p_{X1,X2,X3}(x_1, x_2, x_3) \end{aligned} \quad (91)$$

而 $p_{X3}(x_3)$ 相当于一个向量。

沿着哪个方向求和，就相当于完成了这个维度上数据的合并。这个维度因此便消失。

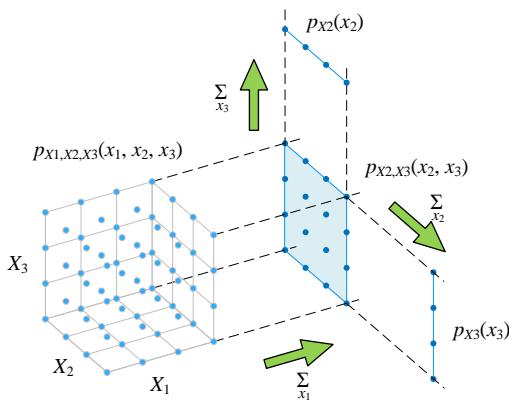


图 48. 先沿 X_1 方向压扁

换个方向， $p_{X2,X3}(x_2, x_3)$ 沿着 x_3 折叠便得到边缘概率质量函数 $p_{X2}(x_2)$ ：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$p_{X_2}(x_2) = \sum_{x_3} p_{X_2, X_3}(x_2, x_3) \quad (92)$$

而 $p_{X_3}(x_3)$ 和 $p_{X_2}(x_2)$ 进一步折叠，便获得概率 1：

$$1 = \sum_{x_3} \sum_{x_2} \sum_{x_1} p_{X_1, X_2, X_3}(x_1, x_2, x_3) = \sum_{x_2} \sum_{x_3} \sum_{x_1} p_{X_1, X_2, X_3}(x_1, x_2, x_3) \quad (93)$$

经过上述不同顺序的三重求和后，三个维度全部消失，结果是样本空间对应的概率值“1”。

请大家沿着上述思路自行分析图 49 两幅图，并写出求和公式。

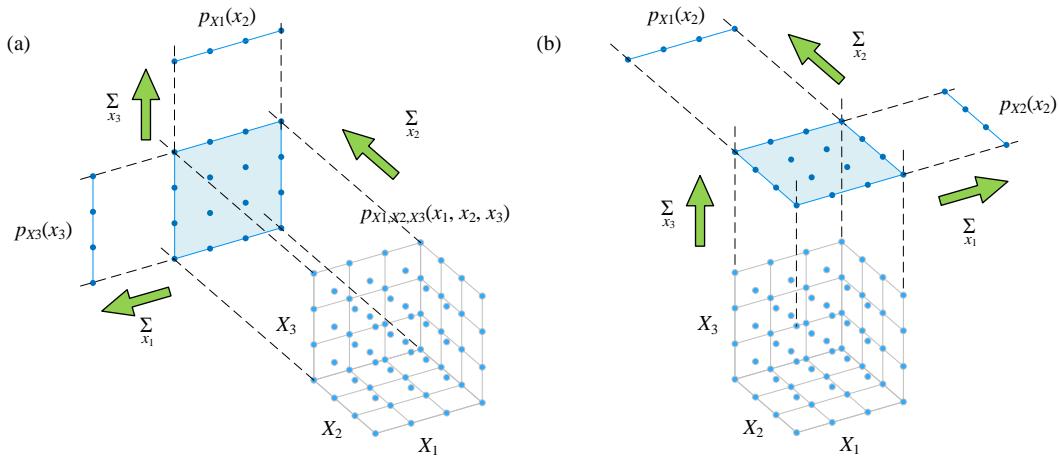


图 49. 分别先沿 X_2 、 X_3 方向压扁

此外，请大家自己思考，如果 X_1 、 X_2 、 X_3 独立，如何计算 $p_{X_1, X_2, X_3}(x_1, x_2, x_3)$ ？

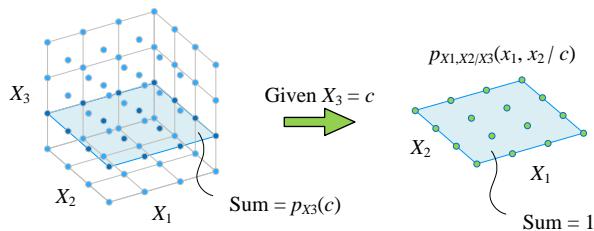
本节 X_1 、 X_2 、 X_3 均为离散随机变量，因此图 48 中每个点均代表概率值。请大家思考以下几种随机变量组合，图 48 这个立方体展开、折叠的方式有何变化？

- ▶ X_1 、 X_2 、 X_3 均为连续随机变量；
- ▶ X_1 、 X_2 为连续随机变量， X_3 为离散随机变量；
- ▶ X_1 、 X_2 为离散随机变量， X_3 为连续随机变量。

条件概率：切片

如图 50 所示，条件概率 $p_{X_1, X_2 | X_3}(x_1, x_2 | c)$ 相当于在 $X_3 = c$ 处切了一片，只考虑切片上的概率分布情况，而不考虑整个立方体的概率分布。

也就是说， $X_3 = c$ 对应的切片是条件概率 $p_{X_1, X_2 | X_3}(x_1, x_2 | c)$ 的样本空间。

图 50. 给定 $X_3 = c$ 条件概率

计算条件概率时，首先将切片上的联合概率求和得到 $p_{X3}(c)$:

$$p_{X3}(c) = \sum_{x_2} \sum_{x_1} p_{X1,X2,X3}(x_1, x_2, c) \quad (94)$$

然后，用联合概率除以 $p_{X3}(c)$ 得到条件概率 $p_{X1,X2|X3}(x_1, x_2 | c)$:

$$p_{X1,X2|X3}(x_1, x_2 | c) = \frac{p_{X1,X2,X3}(x_1, x_2, c)}{p_{X3}(c)} \quad (95)$$

大家自己思考，如果给定 $X_3 = c$ 条件下， X_1 和 X_2 条件独立，意味着什么？



本书第 8 章将继续这个话题。



本章和大家主要探讨了离散随机变量。离散随机变量是指一种在有限或可数的取值集合中随机取值的随机变量。例如，掷硬币的结果只有两个可能的取值：正面或反面，用 0 或 1 来表示。离散随机变量通常用概率质量函数 PMF 来描述其可能取值的概率。对于二元、多元离散随机变量，大家要学会如何计算边缘概率、条件概率。本章最后的鸢尾花例子全面地复盘了有关离散随机变量的关键知识点，请大家务必弄懂。

下一章将介绍离散随机变量中常见分布。

5

Discrete Distributions

离散分布

理想化的离散随机变量概率模型



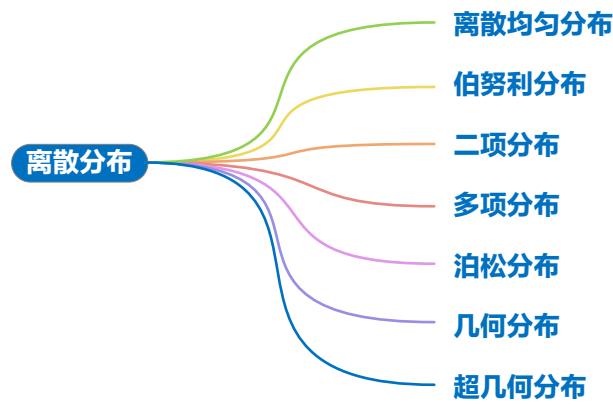
究其本质，概率论无非是将生活常识简化成数学运算。

The theory of probabilities is at bottom nothing but common sense reduced to calculation.

——皮埃尔-西蒙·拉普拉斯 (Pierre-Simon Laplace) | 法国著名天文学家和数学家 | 1749 ~ 1827



- ▶ `matplotlib.pyplot.barh()` 绘制水平直方图
- ▶ `matplotlib.pyplot.stem()` 绘制火柴梗图
- ▶ `mpmath.pi` mpmath 库中的圆周率
- ▶ `numpy.bincount()` 统计列表中元素出现的个数
- ▶ `scipy.stats.bernoulli()` 伯努利分布
- ▶ `scipy.stats.binom()` 二项分布
- ▶ `scipy.stats.geom()` 几何分布
- ▶ `scipy.stats.hypergeom()` 超几何分布
- ▶ `scipy.stats.multinomial()` 多项分布
- ▶ `scipy.stats.poisson()` 泊松分布
- ▶ `scipy.stats.randint()` 离散均匀分布
- ▶ `seaborn.heatmap()` 产生热图



5.1 概率分布：高度理想化的数学模型

本书前文介绍的事件概率描述一次试验中某一个特定样本发生的可能性。想要了解某个随机变量在样本空间中不同样本的概率或概率密度，我们就需要**概率分布** (probability distribution)。

概率分布是一种特殊的函数，它描述随机变量取值的概率规律。概率分布通常包括两个部分：随机变量的取值和对应的概率或概率密度。

和抛物线 $y = ax^2 + bx + c$ 一样，常用的概率分布都是高度理想化的数学模型。

我们知道随机变量分为离散和连续两种，因此概率分布也分为两类——**离散分布** (discrete distribution)、**连续分布** (continuous distribution)。

图 1 给出几种在数据科学、机器学习领域常用的概率分布。图 1 中，用火柴梗图描绘的一元离散随机变量的 PMF，曲线描绘的一元连续随机变量的 PDF。

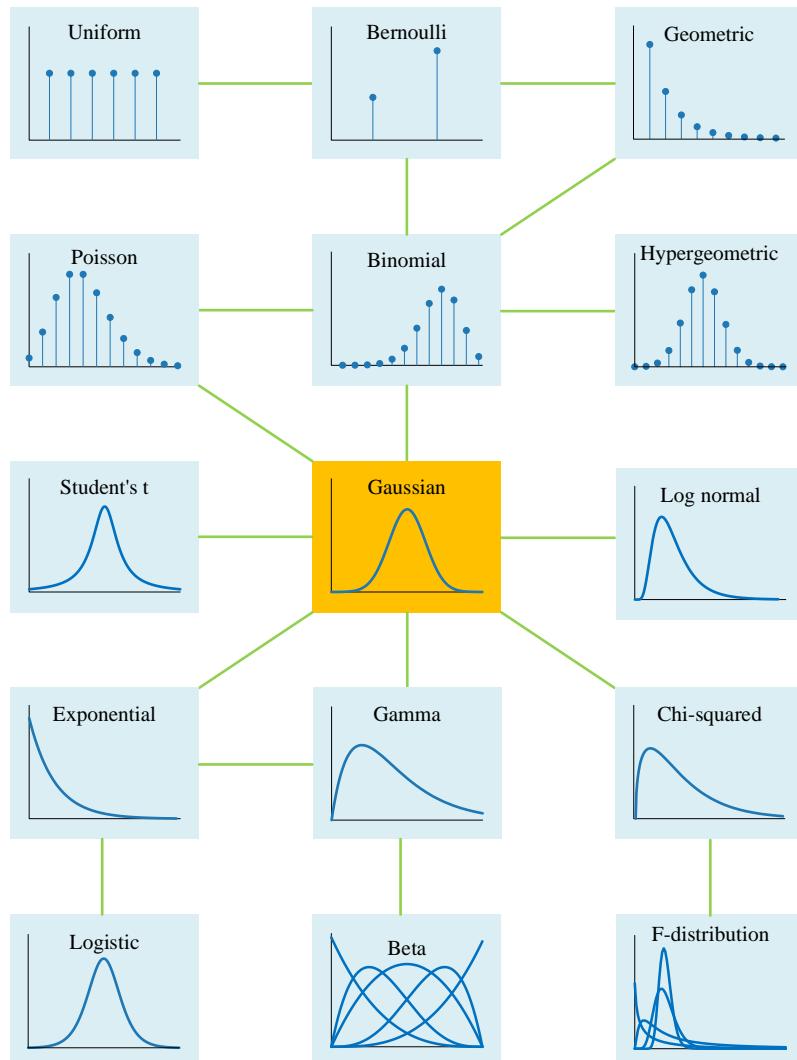


图 1. 常见的几种概率分布，给出多种分布样式

建议大家在学习概率分布时，首先考虑变量是离散还是连续，确定随机变量的取值范围；然后熟悉分布形状以及决定形状的参数，并掌握概率分布的应用场景。

⚠ 再次强调，离散分布对应的是概率质量函数 PMF，其本质是概率。可视化一元、二元离散分布的 PMF 时，建议大家用火柴梗图。连续分布对应的是概率密度函数 PDF。概率密度函数积分、二重积分，有时甚至多重积分后，才得到概率值。可视化一元连续分布 PDF 时，建议用线图，可视化二元连续分布 PDF 时，可以用网格面或等高线。

本章介绍常见离散分布，本书第 7 章讲解连续分布。建议大家把本章和第 7 章当成是“手册”来看待，以浏览的方式来学习，不需要死记硬背各种概率分布函数。大家后续在应用时，如果遇到某个特定概率分布时，可以回来查“手册”。

5.2 离散均匀分布：不分厚薄

离散均匀分布 (discrete uniform distribution) 应该是最简单的离散概率分布。离散型均匀分布分配给离散随机变量所有结果相等的权重。本书前文介绍的抛硬币、抛色子都是离散均匀分布。

离散随机变量 X 等概率地取得 $[a, b]$ 区间内所有整数，取得每一个整数对应的概率为：

$$p_x(x) = \frac{1}{b-a+1}, \quad x = a, a+1, \dots, b-1, b \quad (1)$$

⚠ 注意， a 、 b 为正整数。

上述概率质量函数 $p_x(x)$ 显然满足如下等式：

$$\sum_x p_x(x) = 1 \quad (2)$$

注意，上式是一个函数能够称作一元随机变量 PMF 的基本条件。

期望值、方差

满足 (1) 这个离散均匀分布的 X 的期望值为：

$$E(X) = \frac{a+b}{2} \quad (3)$$

X 的方差为：

$$\text{var}(X) = \frac{(b-a+2)(b-a)}{12} \quad (4)$$

抛骰子试验

定义抛一枚色子结果为离散随机变量 X ，假设获得六个不同点数为等概率，则 X 服从离散均匀分布。 X 的概率质量函数为：

$$p_X(x) = 1/6, \quad x = 1, 2, 3, 4, 5, 6 \quad (5)$$

X 的概率质量函数图像如图 2 所示。请大家自行计算 X 的期望值和方差。

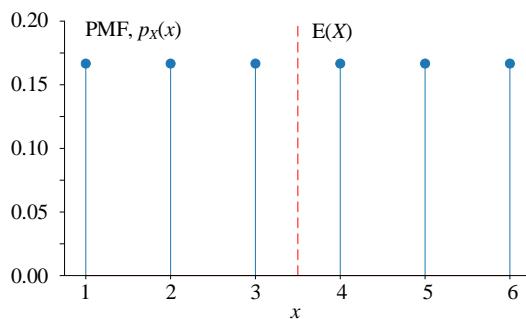
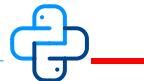


图 2. 离散均匀分布



Bk5_Ch05_01.py 代码文件绘制图 2。

圆周率

我们来看一个《数学要素》第 1 章提过的例子。图 3 所示为圆周率小数点后 1024 位数字的热图。

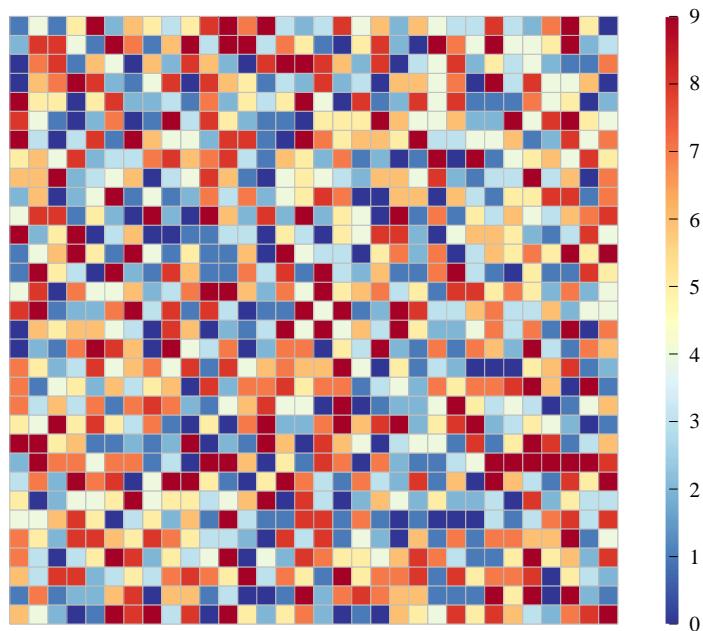


图 3. 圆周率小数点后 1024 位热图，图片来自《数学要素》第 1 章

热图中的数字看似没有任何规律。但是经过分析发现，随着数字数量越大，0 ~ 9 这些数字看上去服从离散均匀分布。图 4 所示为圆周率小数点后 100 位、1,000 位、10,000 位、100,000 位、1,000,000 位 0 ~ 9 这些数字分布。

目前没有关于圆周率是否为 **正规数** (normal number) 的严格证明。正规数是指在某种进位下，其数位上的数字分布均匀、随机且无规律可循的无限小数。具体来说，对于十进制数，每个数字出现的概率应该是相等的，即 $1/10$ 。

尽管圆周率被认为是一种无理数，但它是否为正规数仍然是未解决的问题。在数学上，圆周率和其他著名的无理数，如自然对数的底 e 和 $\sqrt{2}$ ，都被认为是可能是正规数。这些问题是在数学研究中的重要问题，至今仍在继续研究中。

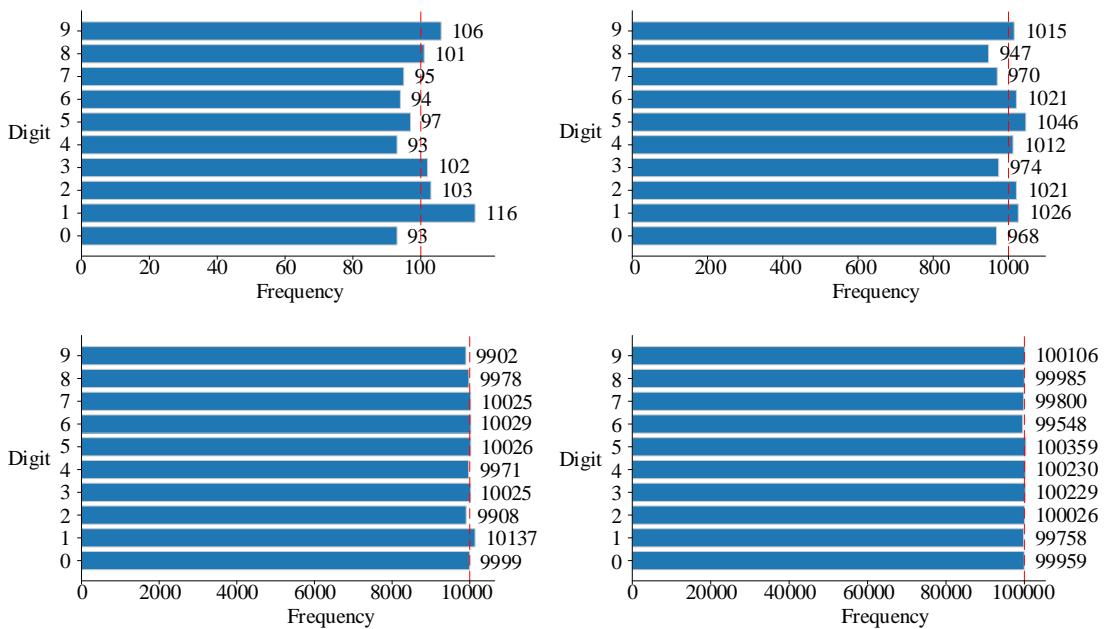


图 4. 圆周率小数点后数字的分布，100 位、1,000 位、10,000 位、100,000 位、1,000,000 位



代码 Bk5_Ch05_02.py 绘制图 3 和图 4。

5.3 伯努利分布：非黑即白

在重复独立试验中，如果每次试验结果离散变量 X 仅有两个可能结果，比如 0、1，这种离散分布叫做**伯努利分布** (Bernoulli distribution)，对应的概率质量函数为：

$$p_x(x) = \begin{cases} p & x=1 \\ 1-p & x=0 \end{cases} \quad (6)$$

其中， p 满足 $0 < p < 1$ 。

(6) 还可以写成：

$$p_x(x) = p^x (1-p)^{1-x} \quad x \in \{0,1\} \quad (7)$$

请大家将 $x=0, 1$ 分别代入上式检验 PMF 结果。

(6) 对应的概率质量函数显然满足归一化条件：

$$\sum_x p_x(x) = p + (1-p) = 1 \quad (8)$$

满足(6)中伯努利分布随机变量 X 的期望和方差分别为：

$$\begin{aligned} E(X) &= p \\ \text{var}(X) &= p(1-p) \end{aligned} \quad (9)$$

抛硬币

本书前文介绍的抛一枚硬币的试验就是常见的伯努利分布。如果硬币质地均匀，获得正面 ($X = 1$)、反面 ($X = 0$) 的概率均为 0.5，则 X 的概率质量函数为：

$$p_X(x) = \begin{cases} 0.5 & x=1 \\ 0.5 & x=0 \end{cases} \quad (10)$$

如果硬币质地不均匀，假设获得正面的概率为 0.6，则对应获得背面的概率为 $1 - 0.6 = 0.4$ 。则 X 的概率质量函数为：

$$p_X(x) = \begin{cases} 0.6 & x=1 \\ 0.4 & x=0 \end{cases} \quad (11)$$

请大家把(10)和(11)写成(7)这种形式。

Python 中伯努利分布函数常用 `scipy.stats.bernoulli()`。

抽样试验

⚠ 再次强调，伯努利分布是离散分布，只有两种对立的可能结果，即结果样本空间只有 2 个元素。伯努利分布的参数只有 p 。

从抽样试验角度，伯努利试验还可以看成是只有两个结果的**放回抽样** (sampling with replacement) 试验。放回抽样中，每次抽样后抽出的样本会被放回总体中，下次抽样时仍然有可能被抽到。与之相对的是**无放回抽样** (sampling without replacement)，在这种情况下，每次抽出的样本不会被放回总体中，下次抽样时不可能再次被抽到。

比如，如图 5 所示，10 只动物中有 6 只兔子、4 只鸡。每次放回抽取一只动物，取到兔子的概率为 0.6，取到鸡的概率为 0.4。

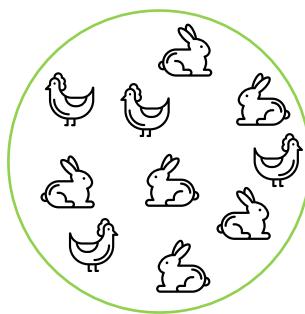


图 5. 从抽样试验角度看伯努利试验

5.4 二项分布：杨辉三角

二项分布 (binomial distribution), 也叫二项式分布, 建立在伯努利分布之上。

举个例子, 一枚硬币抛 n 次, 每次抛掷结果服从伯努利分布, 即正面出现的概率为 p , 反面出现的概率为 $1 - p$, 而且各次抛掷相互独立。进行 n 次独立的试验, 令 X 为获得正面次数, X 对应的概率质量函数:

$$p_X(x) = C_n^x p^x (1-p)^{n-x}, \quad x=0,1,\dots,n \quad (12)$$

(12) 所示二项式概率质量函数 $p_X(x)$ 满足归一化:

$$\begin{aligned} \sum_x p_X(x) &= C_n^0 p^0 (1-p)^n + C_n^1 p^1 (1-p)^{n-1} + \cdots + C_n^n p^n (1-p)^0 \\ &= (p + (1-p))^n = 1 \end{aligned} \quad (13)$$

如果 X 服从 (12) 中给出的二项分布, X 的期望和方差分别为:

$$\begin{aligned} E(X) &= n \cdot p \\ \text{var}(X) &= n \cdot p(1-p) \end{aligned} \quad (14)$$

质地均匀硬币

为了方便大家理解二项分布, 我们假定硬币质地均匀, 即 $p = 0.5$ 。

先从 $n = 1$ 说起, 也就是说试验中抛 1 枚均匀硬币。令 X 为正面为朝上的次数, X 的概率质量函数 PMF 为:

$$p_X(x) = \begin{cases} 1/2 & x=0 \\ 1/2 & x=1 \end{cases} \quad (15)$$

这本质上是伯努利分布。

当 $n = 2$, 即抛 2 枚均匀硬币, X 的概率质量函数为:

$$p_X(x) = \begin{cases} 1/4 & x=0 \\ 1/2 & x=1 \\ 1/4 & x=2 \end{cases} \quad (16)$$

抛 3 枚均匀硬币, X 的概率质量函数为:

$$p_X(x) = \begin{cases} C_3^0 \cdot (1/2)^3 = 1/8 & x=0 \\ C_3^1 \cdot (1/2)^3 = 3/8 & x=1 \\ C_3^2 \cdot (1/2)^3 = 3/8 & x=2 \\ C_3^3 \cdot (1/2)^3 = 1/8 & x=3 \end{cases} \quad (17)$$

试验中, 抛 n 枚均匀硬币, 令 X 为正面为朝上的次数, X 的概率质量函数为:

$$p_X(x) = \begin{cases} C_n^0 \cdot (1/2)^n & x=0 \\ C_n^1 \cdot (1/2)^n & x=1 \\ \dots & \dots \\ C_n^n \cdot (1/2)^n & x=n \end{cases} \quad (18)$$

图 6 所示为 $p = 0.5$ 时， n 取不同值，二项分布的概率质量函数分布。随着 n 不断增大，大家仿佛看到了“高斯分布”。请大家特别注意，高斯分布对应连续随机变量，而二项分布对应离散随机变量。

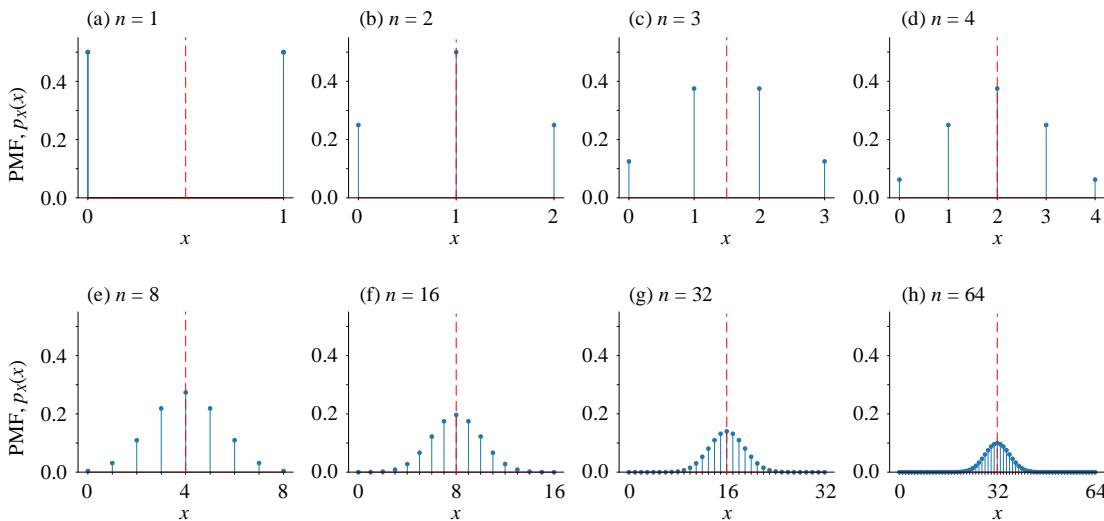


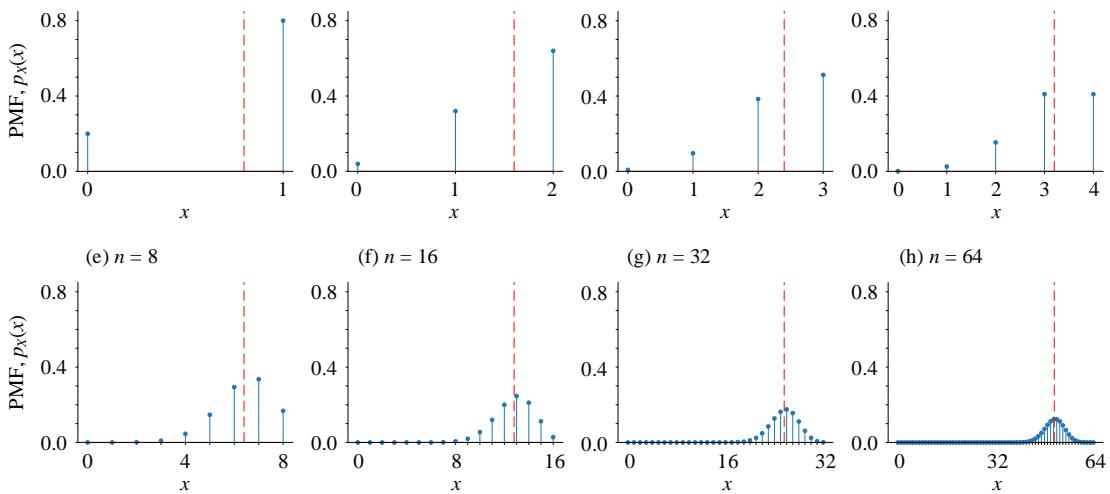
图 6. 二项分布， $p = 0.5$

质地不均匀硬币

如果硬币不均匀，假设正面朝上的概率为 $p = 0.8$ 。试验中，抛硬币 n 次，令 X 为正面朝上的次数，则 X 的概率质量函数为：

$$p_X(x) = \begin{cases} C_n^0 \cdot 0.8^0 (1-0.8)^n & x=0 \\ C_n^1 \cdot 0.8^1 (1-0.8)^{n-1} & x=1 \\ \dots & \dots \\ C_n^n \cdot 0.8^n (1-0.8)^0 & x=n \end{cases} \quad (19)$$

图 7 所示为 $p = 0.8$ 时， n 取不同值，二项分布的概率质量函数分布。

图 7. 二项分布, $p = 0.8$

显然, 二项分布概率质量函数形状是由 n 、 p 两个参数确定下来的。容易发现, 当 $p = 1/2$ 时, PMF 关于 $x = n/2$ 对称。当 $p > 1/2$ 时, PMF 图像偏向 n ; 当 $p < 1/2$ 时, PMF 图像偏向 0。随着 n 不断增大, 分布的偏度逐渐变小, 而且形状上不断近似高斯分布。

⚠ 必须再次强调的是, 二项分布对应离散随机变量, 而高斯分布对应连续随机变量。二项分布 $p_X(x)$ 为概率质量函数, 而高斯分布 $f_X(x)$ 为概率密度函数。

有放回 vs 不放回

总结来说, 二项分布是 n 个独立进行的伯努利试验。二项分布 PMF 有两个参数—— n 、 p 。

从抽样试验角度, 二项分布强调“独立”, 每次抽取后再放回, 这样总体本身不发生变化。还是利用鸡兔做例子, 每次抽取时, 取得兔子的概率为 0.6, 取得鸡的概率为 0.4。计算 $n = 10$ 次有放回抽取中有 5 只兔子的概率, 用的就是二项分布。

若是不放回抽样, 即每次抽样之后不放回, 则总体随之变化, 分别取得鸡、兔的概率不断变化。二项分布则无法处理无放回抽样, 我们需要用到超几何分布。超几何分布是本章后续要介绍的分布类型。

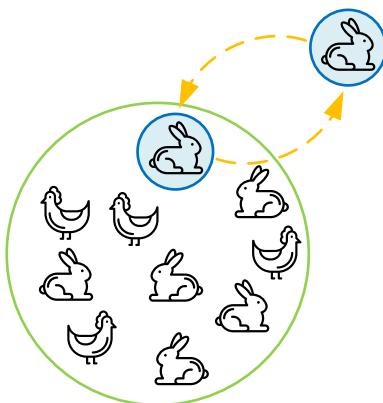


图 8. 从抽样试验角度看二项分布



代码 Bk5_Ch05_03.py 绘制图 6 和图 7。

5.4 多项分布：二项分布推广

多项分布 (multinomial distribution)，也叫多项式分布，是二项式分布的推广。多项分布描述在 n 次独立重复的试验中，每次试验有 K 个可能的结果中的一个发生的次数的概率分布。每次试验的 K 个可能结果的概率不一定相等。

多项分布的概率质量函数为：

$$p_{x_1, \dots, x_K}(x_1, \dots, x_K; n, p_1, \dots, p_K) = \begin{cases} \frac{n!}{(x_1!) \times (x_2!) \cdots \times (x_K!)} \times p_1^{x_1} \times \cdots \times p_K^{x_K} & \text{when } \sum_{i=1}^K x_i = n \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

其中 $x_i (i = 1, 2, \dots, K)$ 为非负整数，且 $\sum_{i=1}^K p_i = 1$ 。这个分布常记做 Mult(p) 或 Mult(p_1, p_2, \dots, p_K)。

⚠ 注意，为了避免混淆，本书用 “|” 引出条件概率中的条件，用分号 “;” 引出概率分布的参数。

特别地，如果 $n = 1$ ，多项分布就变成了**类别分布** (categorical distribution)。

举个例子

假设一个农场有大量动物，其中 60% 为兔子，10% 为猪，30% 为鸡。如果随机抓取 8 只动物，其中有 2 只兔子、3 只猪、3 只鸡的概率为多少？

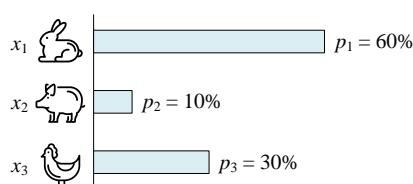


图 9. 农场兔、猪、鸡的比例

计算这个概率就用到了多项分布。当 $K = 3$ 且 $n = 8$ 时，多项式分布的概率质量函数为：

$$f(x_1, x_2, x_3; p_1, p_2, p_3) = \begin{cases} \frac{8!}{(x_1!) \times (x_2!) \times (x_3!)} \times p_1^{x_1} \times p_2^{x_2} \times p_3^{x_3} & \text{when } x_1 + x_2 + x_3 = 8 \\ 0 & \text{otherwise} \end{cases} \quad (21)$$

其中， x_1 、 x_2 、 x_3 均为非负整数。

将 $x_1 = 2$, $x_2 = 3$, $x_3 = 3$, $p_1 = 0.6$, $p_2 = 0.1$, $p_3 = 0.3$ 代入上式得到：

$$f\left(\begin{array}{ccccc} 2, 3, 3; & 0.6, 0.1, 0.3 \\ x_1 & x_2 & x_3 & p_1 & p_2 & p_3 \end{array}\right) = \frac{8!}{(2!) \times (3!) \times (3!)} \times 0.6^2 \times 0.1^3 \times 0.3^3 \approx 0.0054 \quad (22)$$

散点图、热图、火柴梗图

下面，我们分别用三维散点图、二维散点图、热图、火柴梗图可视化多项分布。

给定参数， $n = 8$, $p_1 = 0.6$, $p_2 = 0.1$, $p_3 = 0.3$ ，多项分布的三维散点图如图 10 (a) 所示。图中每一个散点代表一个 (x_1, x_2, x_3) 组合，注意这三个均为非负整数。由于 $x_1 + x_2 + x_3 = 8$ ，所以 (x_1, x_2, x_3) 散点均在一个平面上。散点的颜色代表概率质量 PMF 值大小。

将这些散点投影在 $x_1 x_2$ 平面上，便得到图 10 (b)。这说明只要给定 x_1 和 x_2 ，根据 $x_3 = 8 - (x_1 + x_2)$ ， x_3 便确定下来。

图 11 所示为上述多项分布的 PMF 热图和散点图。

图 12、图 13 和图 14、图 15 可可视化另外两组参数的多项分布，请大家自行比较分析。

→ 二项分布、多项分布、Beta 分布、Dirichlet 分布（第 7 章）经常一起出现在贝叶斯推断（Bayesian inference）中，这是本书第 21、22 章要介绍的内容。

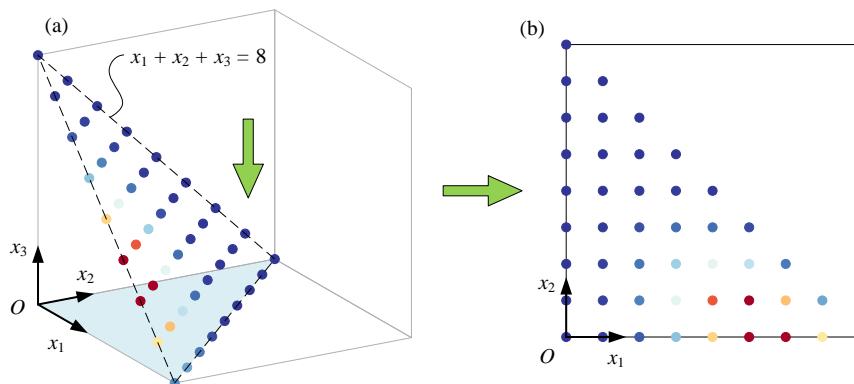
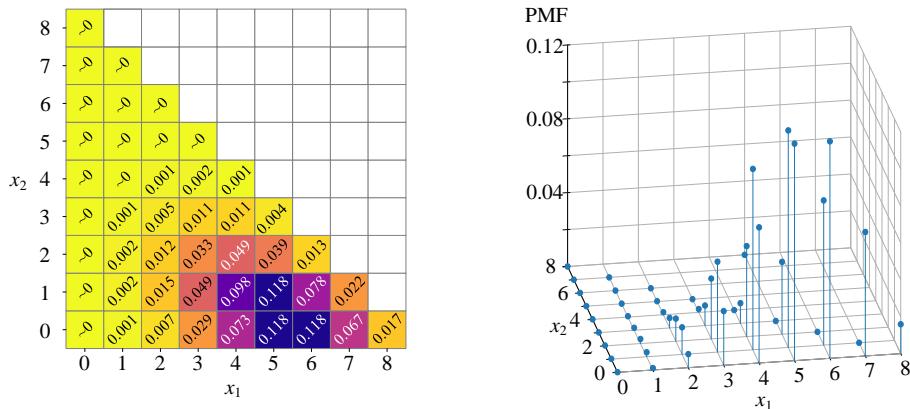
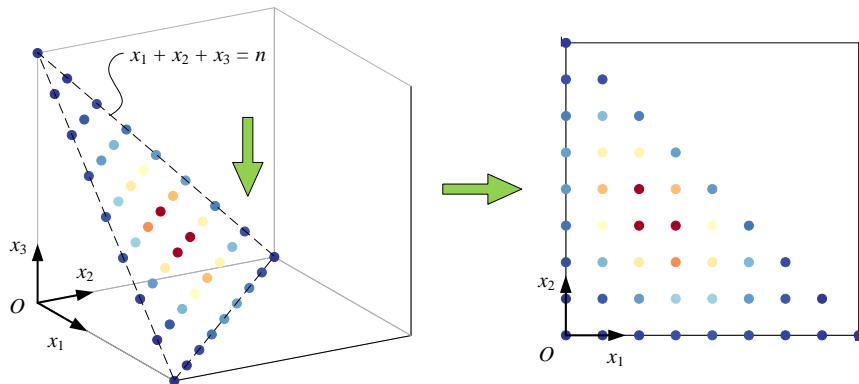
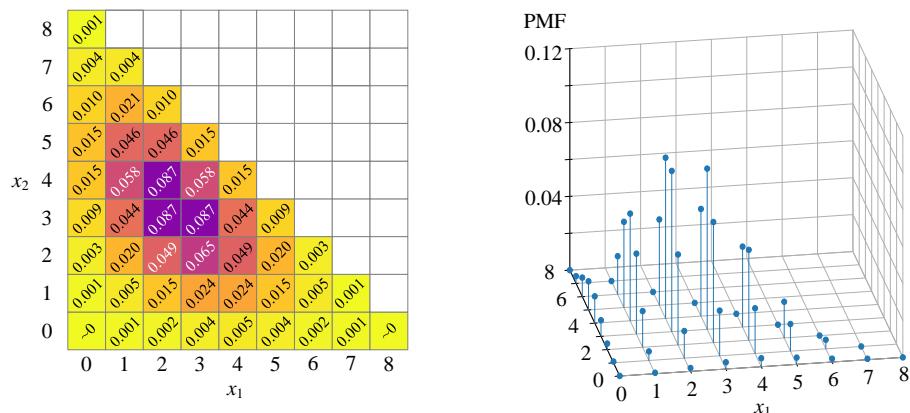


图 10. 多项分布 PMF 三维和平面散点图, $n = 8$, $p_1 = 0.6$, $p_2 = 0.1$, $p_3 = 0.3$

图 11. 多项分布 PMF 热图和火柴梗图, $n = 8$, $p_1 = 0.6$, $p_2 = 0.1$, $p_3 = 0.3$ 图 12. 多项分布 PMF 三维和平面散点图, $n = 8$, $p_1 = 0.3$, $p_2 = 0.4$, $p_3 = 0.3$ 图 13. 多项分布 PMF 热图和火柴梗图, $n = 8$, $p_1 = 0.3$, $p_2 = 0.4$, $p_3 = 0.3$

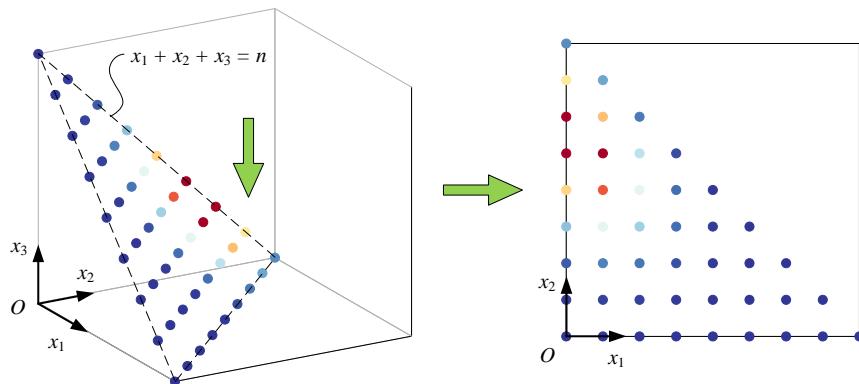
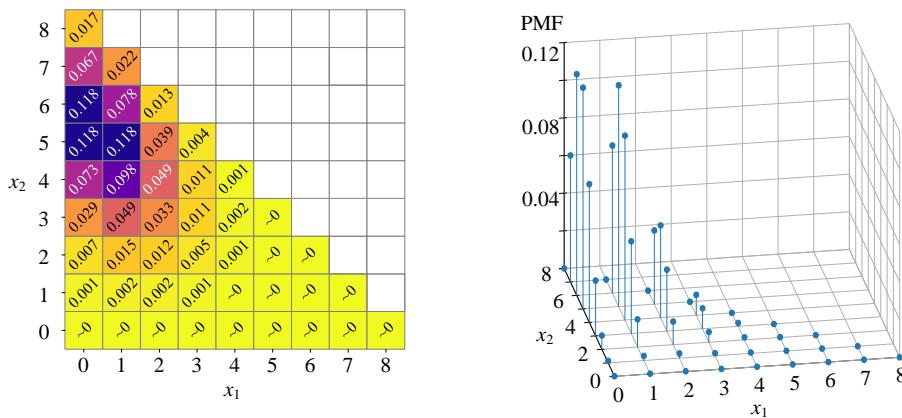
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 14. 多项分布 PMF 三维和平面散点图, $n = 8$, $p_1 = 0.1$, $p_2 = 0.6$, $p_3 = 0.3$ 图 15. 多项分布 PMF 热图和火柴梗图, $n = 8$, $p_1 = 0.1$, $p_2 = 0.6$, $p_3 = 0.3$ 

Bk5_Ch05_04.py 绘制本节图像。

5.5 泊松分布：建模随机事件的发生次数

如果二项分布的试验次数 n 非常大，事件每次发生的概率 p 非常小，并且它们的乘积 np 存在有限的极限 λ ，则这个二项分布趋近于另一种分布——**泊松分布** (Poisson distribution)。泊松分布是一种离散型概率分布，它描述的是在一定时间内某个事件发生的次数。

泊松分布的概率质量函数为：

$$p_x(x) = \frac{\exp(-\lambda)\lambda^x}{x!}, \quad x=0,1,2,\dots \quad (23)$$

图 16 所示为泊松分布概率质量函数随 λ 变化。

满足 (23) 泊松随机变量的期望和方差都是 λ :

$$\mathbb{E}(X) = \text{var}(X) = \lambda \quad (24)$$

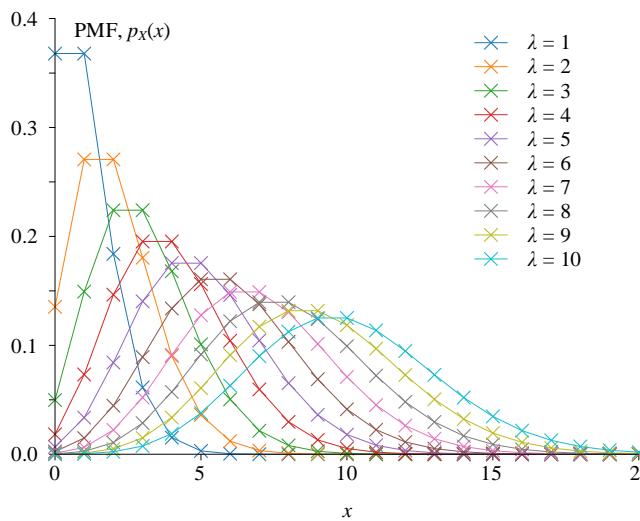


图 16. 泊松分布概率质量函数随 λ 变化

我们一般用泊松分布描述在给定的时间段、距离、面积等范围内随机事件发生的概率。应用泊松分布的例子包括每小时走入商店的人数，一定时间内机器出现故障的次数，一定时间内交通事故发生的次数等等。

⚠ 再次强调一下，泊松分布的均值和方差相等，都等于 λ 。这也就意味着，当 λ 确定时，泊松分布的形态也就确定了。



代码 Bk5_Ch05_05.py 绘制图 16。

5.6 几何分布：滴水穿石

几何分布 (geometric distribution) 也是一个单参数概率分布，几何分布模拟一系列独立伯努利试验中一次成功之前的失败次数。其中，每次试验要么成功要么失败，并且任何单独试验的成功概率是恒定的。

比如，抛 x 次硬币 (伯努利试验)，前 $x - 1$ 次均为反面，在第 x 次为正面。

在连续抛硬币的试验中，每次抛掷正面出现的概率为 p ，反面出现的概率为 $1 - p$ ，每次抛掷相互独立。令 X 为连续抛掷一枚硬币，直到第一次出现正面所需要的次数。 X 的概率质量函数为：

$$p_x(x) = (1 - p)^{x-1} p, \quad x = 1, 2, \dots \quad (25)$$

满足 (25) 几何分布的离散随机变量 X 的期望和方差分别为：

$$\begin{aligned} E(X) &= \frac{1}{p} \\ \text{var}(X) &= \frac{1-p}{p^2} \end{aligned} \quad (26)$$

图 17 所示为当 $p = 0.5$ 时，几何分布概率质量函数 PMF 和 CDF。

注意，几何分布的随机变量有两种定义：1) 获得一次成功所需要的试验次数；2) 第一成功之前经济的失败次数。两者之差为 1。它们的期望值也不同。

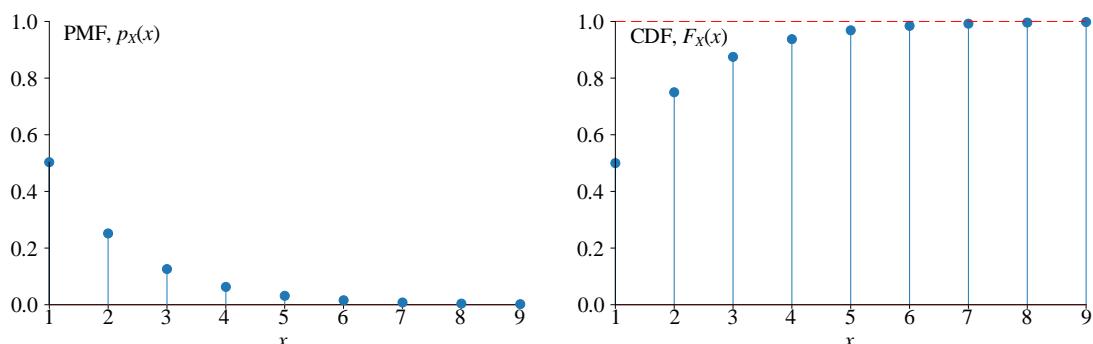
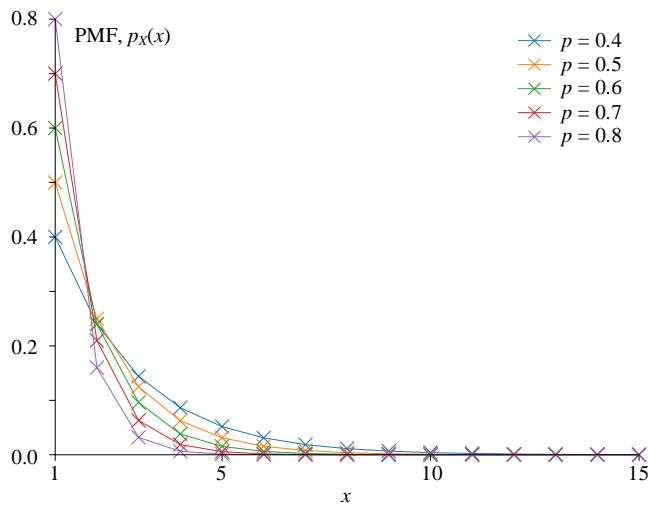
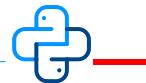


图 17. 几何分布概率质量函数 PMF 和 CDF, $p = 0.5$

图 18 所示为几何分布概率质量函数 PMF 随 p 变化。

图 18. 几何分布概率质量函数 PMF 随 p 变化

代码 Bk5_Ch05_06.py 绘制图 17 和图 18。

5.7 超几何分布：不放回

我们在介绍二项分布时，特别强调二项分布在抽样时放回。如果抽样时不放回，我们便得到超几何分布 (hypergeometric distribution)。

举个例子，假如某个农场总共有 N 个动物，其中 K 只兔子。从 N 只动物不放回抽取 n 个动物，其中有 x 只兔子的概率为：

$$p_X(x) = \frac{C_K^x C_{N-K}^{n-x}}{C_N^n}, \quad \max(0, n+K-N) \leq x \leq \min(K, n) \quad (27)$$

这个分布就是超几何分布。

比如，如图 19 所示，有 50 (N) 只动物，其中有 15 (K) 只兔子 (30%)。从 50 (N) 只动物中不放回地抽取 20 (n) 只动物，其中有 x 只兔子对应的概率为：

$$p_X(x) = \frac{C_{15}^x C_{35}^{20-x}}{C_{50}^{20}} \quad (28)$$

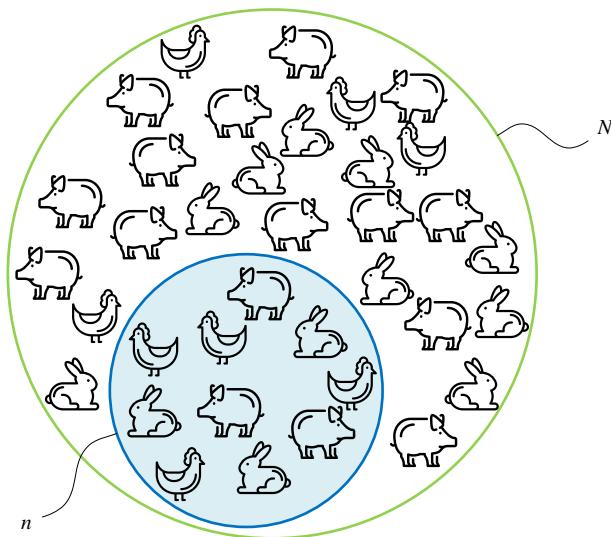
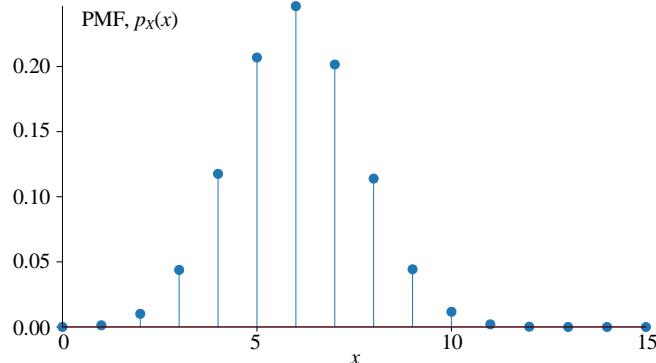
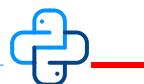


图 19. 超几何分布原理

上式中概率质量函数 $p_X(x)$ 对应的图像如图 20 所示。

总结来说，超几何分布的核心是“不放回”。超几何分布 PMF 输入有四个， N 、 K 描述整体， n 、 x 描述采样。

图 20. 超几何分布概率质量函数， $N = 50$, $K = 15$, $n = 20$ 

代码 Bk5_Ch05_07.py 绘制图 20。

二项分布 vs 超几何分布

如果总体数量 N 很大，抽取数量 n 很小，不管抽样时是否放回，都可以用二项分布近似。

举个例子，兔子占整体的比例确定为 $p = 0.3$ (30%)，而动物总体数量分别为 $N = 100, 200, 400, 800$ 条件下，放回抽取 (二项分布)、不放回抽取 (超几何分布) $n = 20$ 只动物，兔子数量 x 对应概率分布如图 21 所示。

观察这四幅子图，我们发现当 N 不断增大，二项分布和超几何分布的 PMF 曲线逐渐靠近。

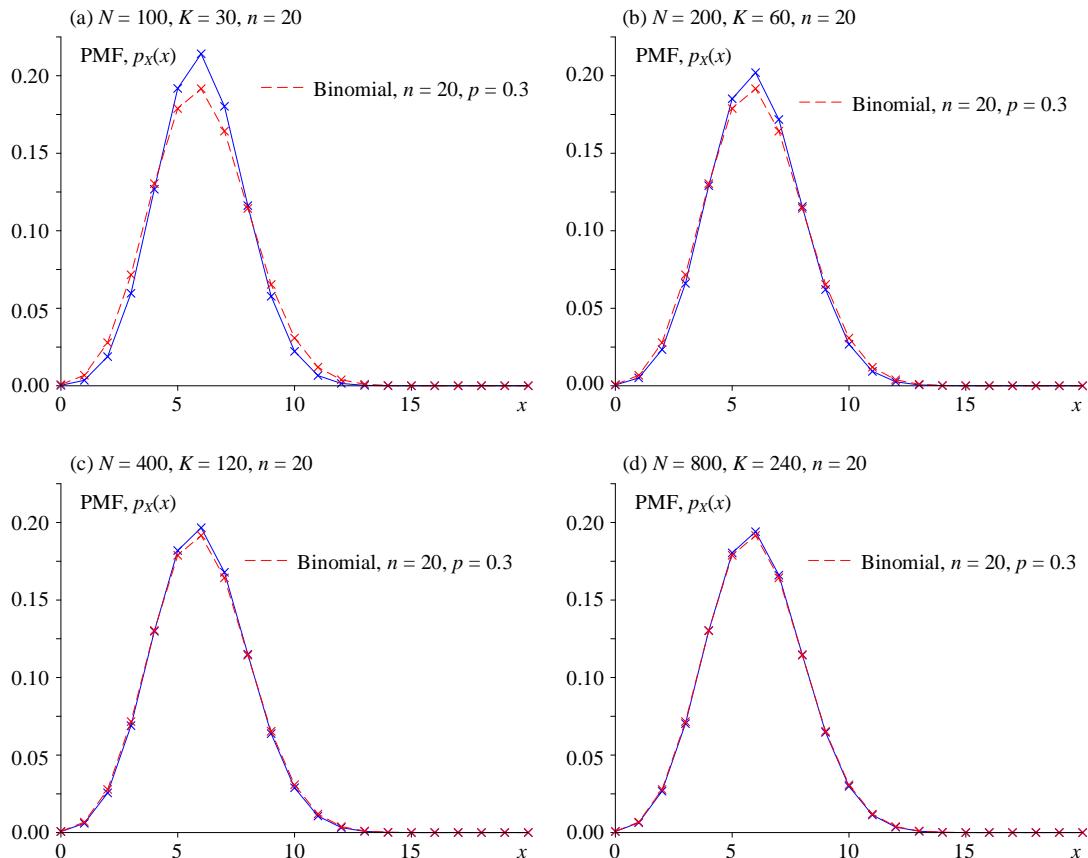
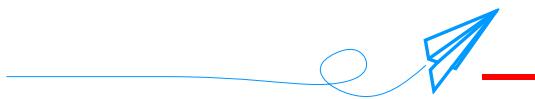


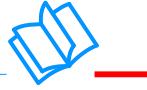
图 21. 超几何分布 PMF 和二项分布 PMF 关系



代码 Bk5_Ch05_08.py 绘制图 21。



离散分布是概率论中的一种重要分布类型，描述的是在一定条件下，随机变量取值的概率分布情况。离散分布也是高度理想化的数学模型，一种近似而已。这一章需要大家格外留意二项分布、多项分布，它们在本书贝叶斯推断中将起到重要作用。



各种分布之间的联系，请大家参考：

<http://www.math.wm.edu/~leemis/chart/UDR/UDR.html>

6

Continuous Random Variables

连续随机变量

PDF 积分得到边缘概率密度或概率值



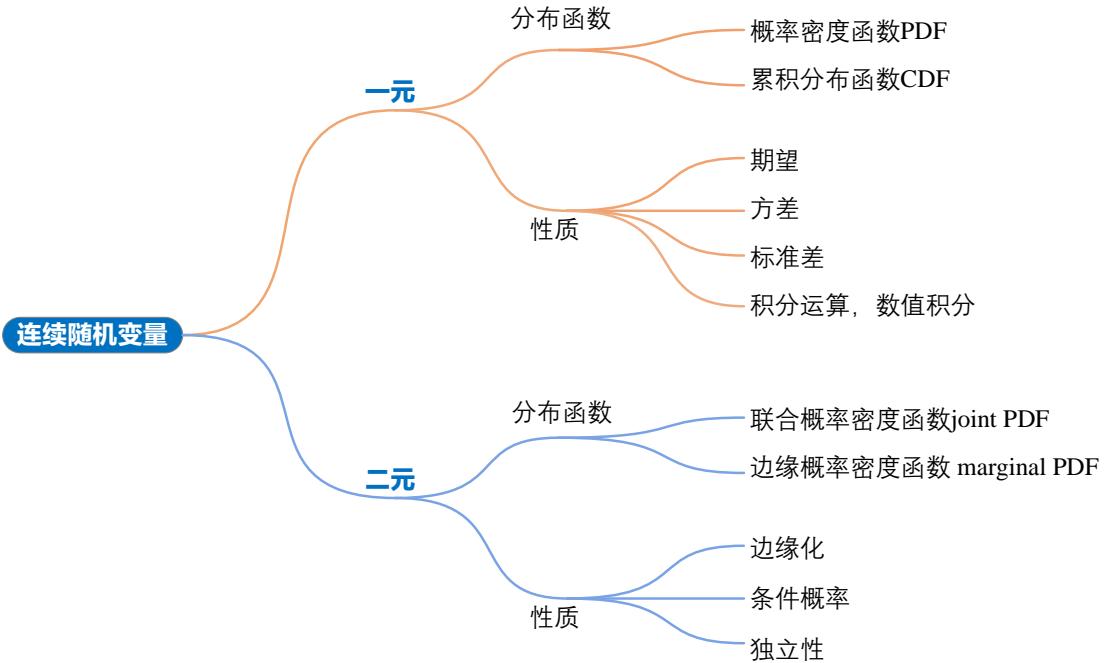
上帝不仅玩骰子，他还有时把骰子扔到人类看不见的地方。

Not only does God definitely play dice, but He sometimes confuses us by throwing them where they can't be seen.

—— 史蒂芬·霍金 (Stephen Hawking) | 英国理论物理学家、宇宙学家 | 1942 ~ 2018



- ◀ `matplotlib.pyplot.contour()` 绘制平面等高线
- ◀ `matplotlib.pyplot.contour3D()` 绘制三维等高线
- ◀ `matplotlib.pyplot.contourf()` 绘制平面填充等高线
- ◀ `matplotlib.pyplot.fill_between()` 区域填充颜色
- ◀ `matplotlib.pyplot.plot_wireframe()` 绘制三维单色线框图
- ◀ `matplotlib.pyplot.scatter()` 绘制散点图
- ◀ `scipy.stats.st.gaussian_kde()` 高斯 KDE 函数
- ◀ `seaborn.scatterplot()` 绘制散点图
- ◀ `statsmodels.api.nonparametric.KDEUnivariate()` 一元核密度估计



6.1 一元连续随机变量

本书第4章区分过**离散随机变量**(discrete random variable)、**连续随机变量**(continuous random variable)。如果随机变量 X 的所有可能取值不可以逐个列举出来，而是整个数轴或数轴上某一区间内的任一点，我们就称 X 为连续随机变量。

概率密度函数：积分

本书第4章介绍过，离散随机变量对应的数学工具为求和 Σ ，连续随机变量对应积分 \int 。对于连续随机变量 X ，如果存在非负函数 $f_X(x)$ 使得：

$$\Pr(X \in B) = \int_B f_X(x) dx \quad (1)$$

则称函数 $f_X(x)$ 为 X 的**概率密度函数**(probability density function, PDF)。

特别地，如图1所示，当 B 为区间 $[a, b]$ 时，随机变量 X 的概率对应定积分：

$$\Pr(a \leq X \leq b) = \int_a^b f_X(x) dx \quad (2)$$

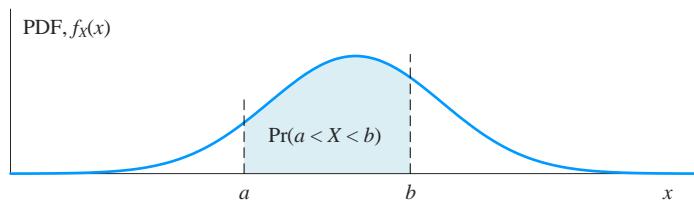


图1. 定积分常用来计算一元连续随机变量在一定区间对应的概率

此外，本书前文提到过，PMF和PDF的输入都可能是不止一个随机变量，这和多元函数一样。比如，二元连续随机变量 (X, Y) 联合概率密度函数PDF $f_{X,Y}(x,y)$ 有两个变量，三元连续随机变量 (X_1, X_2, X_3) 的联合概率密度函数PDF $f_{X_1,X_2,X_3}(x_1,x_2,x_3)$ 有三个变量。

概率密度非负，面积为1

概率密度函数 $f_X(x)$ 必须是非负 $f_X(x) \geq 0$ ，且满足：

$$\Pr(-\infty < X < \infty) = \int_{-\infty}^{\infty} f_X(x) dx = 1 \quad (3)$$

上式常简写为：

$$\int_x f_X(x) dx = 1 \quad (4)$$

如图 2 所示，从图像上来看， $f_X(x)$ 曲线和整个横轴包围区域的面积为 1，这也是归一化。换句说话，一个函数要想能当做概率密度函数来用先要满足非负、面积为 1 这两个条件。

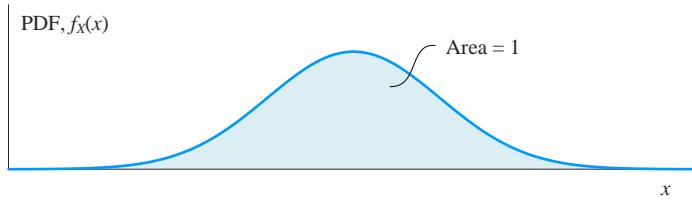


图 2. $f_X(x)$ 和横轴围成图形的面积为 1

单点集合：概率密度非负，但是概率为 0

利用数值积分方法， X 的取值范围在 $[a, a + \Delta]$ 对应的概率为：

$$\Pr(a \leq X \leq a + \Delta) = \int_a^{a+\Delta} f_X(x) dx \approx f_X(a)\Delta \quad (5)$$

当 $\Delta \rightarrow 0$ 时， $\Pr(a \leq X \leq a + \Delta) \rightarrow 0$ 。

也就是说，对于单点集合， $X = a$ 的概率为 0：

$$\Pr(X = a) = \int_a^a f_X(x) dx = 0 \quad (6)$$

即便概率密度 $f_X(a)$ 大于 0。

区间端点

因此，对于连续随机变量 X ，区间端点对概率计算不起任何作用，因此以下四个概率值等价：

$$\Pr(a \leq X \leq b) = \Pr(a < X \leq b) = \Pr(a \leq X < b) = \Pr(a < X < b) \quad (7)$$

这就好比“单丝不成线、独木不成林”。这一点，连续随机变量、离散随机变量完全不同。

概率密度值可以大于 1

再次强调 $f_X(x)$ 并不是概率，而是概率密度，因此 $f_X(x)$ 可大于 1。

比如，图 3 所示在 $[0, 0.5]$ 区间上连续均匀分布的概率密度函数 $f_X(x)$ 。很明显， $f_X(x)$ 的最大值为 2，但是长方形的面积仍为 1：

$$\begin{aligned}\Pr(-\infty < X < \infty) &= \int_{-\infty}^0 f_X(x) dx + \int_0^{0.5} f_X(x) dx + \int_{0.5}^{\infty} f_X(x) dx \\ &= 0 + \int_0^{0.5} 2 dx + 0 \\ &= 2x \Big|_0^{0.5} = 1\end{aligned}\tag{8}$$

⚠ 反复强调，图 3 中的 2 不是概率值，而是概率密度。对于一元随机变量，概率密度函数在一定区间内积分结果才是概率值。概率密度虽然不是概率值，但也量化“可能性”。

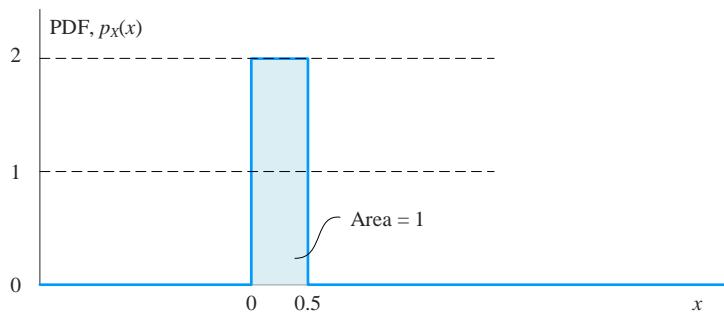


图 3. 概率密度函数 $f_X(x)$ 可以大于 1

累积分布函数

本书前文介绍，给定一元离散随机变量 X 的概率质量函数 $p_X(x)$ ，求解其 CDF 时，用的是累加 Σ 。

以图 4 (a) 为例，对于一元连续随机变量 X ，求累积分布函数 CDF $F_X(x)$ 用的是积分，也就是求面积：

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f_X(t) dt\tag{9}$$

图 4 (a) 中 $f_X(x)$ 图形的面积对应概率值，而图 4 (b) 中 $F_X(x)$ 的高度对应概率值。

随机变量 X 在 $[a, b]$ 区间对应的概率可以用 CDF $F_X(x)$ 计算：

$$\Pr(a \leq X \leq b) = F_X(b) - F_X(a)\tag{10}$$

再次强调，对于一元连续随机变量，PDF 是概率密度，CDF 是概率。

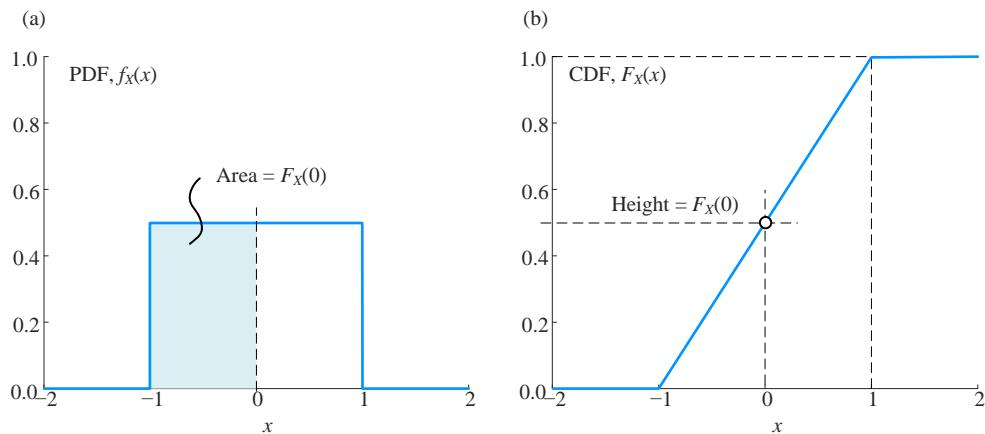


图 4. 连续均匀分布 PDF 和 CDF

6.2 期望、方差和标准差

期望值

连续随机变量 X 期望定义如下：

$$E(X) = \int_{-\infty}^{\infty} x \cdot \underbrace{f_X(x)}_{\text{Weight}} dx \quad (11)$$

上式也相当于加权平均。其中， $f_X(x)$ 相当于“权重”。显然， $f_X(x)$ 非负，但是 x 取值可正可负。这也就是说， $E(X)$ 可正可负。

(11) 常简写为：

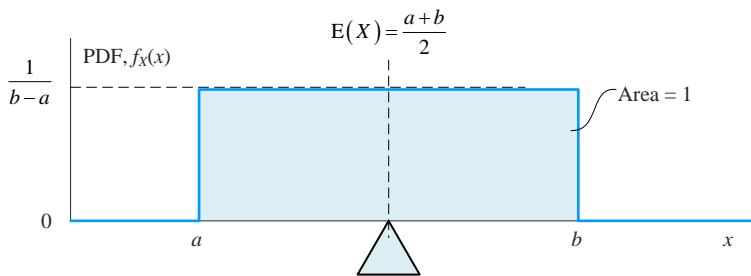
$$E(X) = \int_x x \cdot f_X(x) dx \quad (12)$$

权重当然满足 $\int_x f_X(x) dx = 1$ 。

连续均匀分布

如图 5 所示，如果随机变量 X 在 $[a, b]$ 上服从 **连续均匀分布** (continuous uniform distribution)， X 的概率密度函数为：

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \quad (13)$$

图 5. 随机变量 X 在 $[a, b]$ 上为均匀分布

X 的期望值为：

$$\mathbb{E}(X) = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{1}{b-a} \frac{b^2 - a^2}{2} = \frac{a+b}{2} \quad (14)$$

随机变量 X 的取值在 $[a, b]$ 变化，对应的概率密度变化用 $f_X(x)$ 刻画。而求得的期望值 $\mathbb{E}(X)$ 则是一个标量，这相当于总结归纳，也是降维。

几何角度来看，如图 5 所示，计算 X 的期望值相当于找到一块均质木板的质心在长度方向上的位置。



相比于第 4 章的离散随机变量求和运算，积分运算可以看做是“极尽细腻”的求和。

方差

连续随机变量 X 方差的定义为：

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \int_x \underbrace{(x - \mathbb{E}(X))^2}_{\text{Deviation}} \cdot \underbrace{f_X(x)}_{\text{Weight}} dx \quad (15)$$

同样，连续随机变量 X 的方差也满足如下计算技巧：

$$\text{var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \quad (16)$$

其中，

$$\mathbb{E}(X^2) = \int_x x^2 \cdot f_X(x) dx \quad (17)$$

举个例子

对于图 5 所示均匀分布，为了方便计算 X 的方差，计算 X 平方的期望值为：

$$\mathbb{E}(X^2) = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^3}{3} \Big|_a^b = \frac{1}{b-a} \frac{b^3 - a^3}{3} = \frac{a^2 + ab + b^2}{3} \quad (18)$$

根据(16), X 的方差为：

$$\begin{aligned} \text{var}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12} \end{aligned} \quad (19)$$

数值积分

如图 6 所示，随机变量 X 在 $[0, 1]$ 上为均匀分布。我们可以很容易通过积分得到期望值、方差。但是，并不是所有的概率密度函数都有解析式；此外，即便概率密度函数有解析式，也不代表我们能计算得到积分的解析解，比如高斯函数。

如图 7 所示，这就需要用到《数学要素》第 18 章介绍的**数值积分** (numerical integration)。当然，我们还可以用**蒙特卡洛模拟** (Monte Carlo simulation) 估算面积，这是本书后续要介绍的内容。

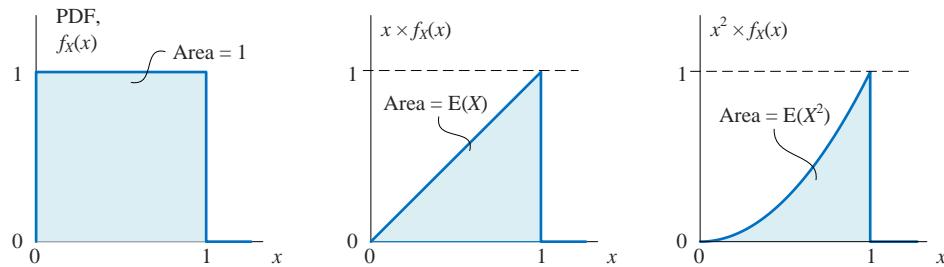


图 6. 随机变量 X 在 $[0, 1]$ 上为均匀分布

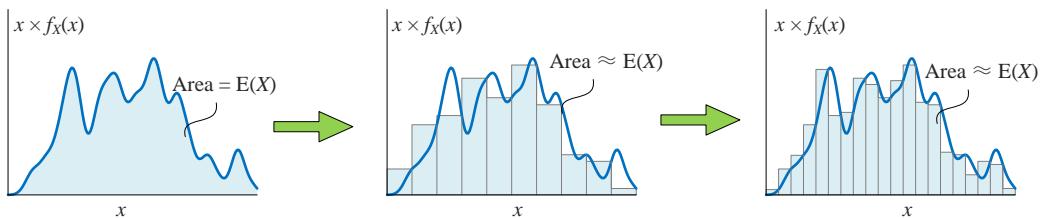


图 7. 数值积分估算期望值

6.3 二元连续随机变量

假设同一个试验中，有两个连续随机变量 X 和 Y ，非负二元函数 $f_{X,Y}(x,y)$ 为 (X, Y) 的**联合概率密度函数** (joint probability density function 或 joint PDF)。

本章前文介绍，对于一元连续随机变量，积分得到的面积对应概率。而二元随机变量计算概率的工具是二重积分，从图像上来看，二重积分得到的体积对应概率。

如图 8 所示，给定积分区域 $A = \{(x, y) | a < x < b, c < y < d\}$ ，概率 $\Pr((X, Y) \in A)$ 对应的二重积分为：

$$\underbrace{\Pr((X, Y) \in A)}_{\text{Probability}} = \int_c^d \int_a^b f_{X,Y}(x, y) dx dy \quad (20)$$

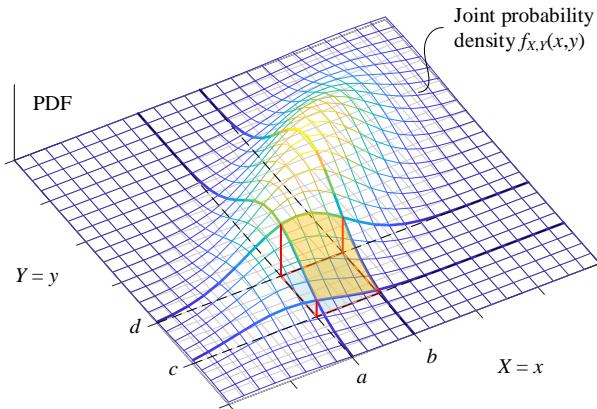


图 8. 二元 PDF $f_{X,Y}(x,y)$ 在 $A = \{(x, y) | a < x < b, c < y < d\}$ 二重积分

体积为 1：样本空间概率为 1

如果积分区域为整个平面，二重积分的结果为 1：

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{X,Y}(x, y) dx dy = 1 \quad (21)$$

也就是说，图 8 中 $f_{X,Y}(x,y)$ 曲面和水平面围成几何形状的体积为 1，代表样本空间的概率为 1。上式本质上也是“穷举法”。

累积概率密度 CDF

二元累积概率函数 CDF $F_{X,Y}(x,y)$ 定义为：

$$\underbrace{F_{X,Y}(x, y)}_{\text{Probability}} = \Pr(X < x, Y < y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(s, t) ds dt \quad (22)$$

图 9 所示等高线为某个二元累积概率函数 $F_{X,Y}(x,y)$ 。图 9 还绘制了两条边缘 CDF 曲线。

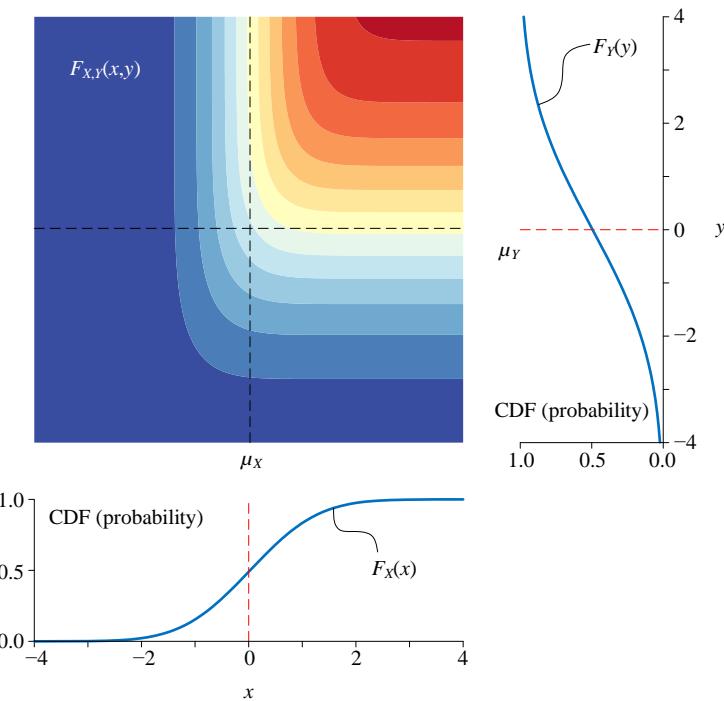


图 9. CDF 函数曲面 $F_{X,Y}(x,y)$ 平面填充等高线，边缘 CDF

6.4 边缘概率：二元 PDF 偏积分

图 10 所示为二元概率密度函数 $f_{X,Y}(x,y)$ 曲面和边缘概率密度曲线的关系。

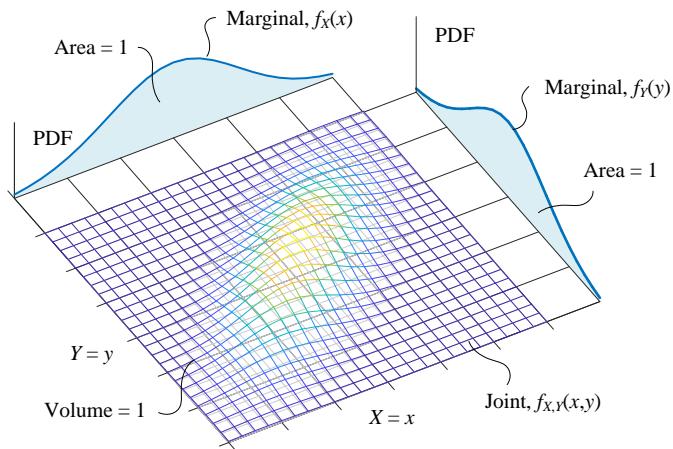


图 10. 二元联合概率密度函数曲面和边缘概率密度之间的关系

边缘概率密度函数 $f_X(x)$

如图 11 所示，连续随机变量 X 的边缘概率密度函数 $f_X(x)$ 可以通过 $f_{X,Y}(x,y)$ 对 y “偏积分”得到：

$$\underbrace{f_X(x)}_{\text{Marginal}} = \overbrace{\int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy}^{\substack{\text{Eliminate } y \\ \text{Joint}}} \quad (23)$$

上式，相当于消去（降维、压扁、折叠）变量 y ，这和离散随机变量的“偏求和”类似。

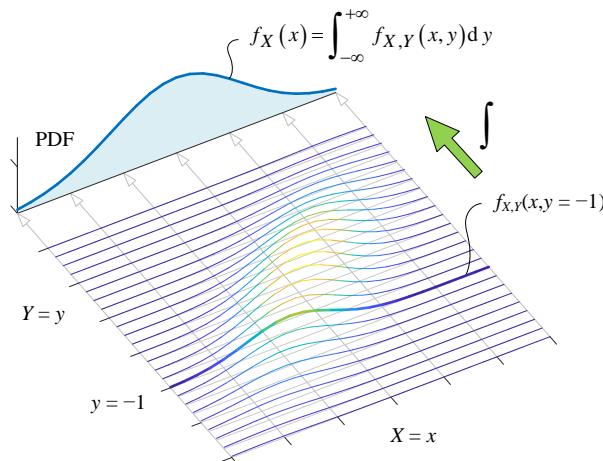


图 11. 联合概率密度 $f_{X,Y}(x,y)$ 对 y “偏积分”得到边缘概率密度 $f_X(x)$

(23) 可以简写为：

$$\underbrace{f_X(x)}_{\text{Marginal}} = \overbrace{\int_y f_{X,Y}(x,y) dy}^{\substack{\text{Eliminate } y \\ \text{Joint}}} \quad (24)$$

⚠ 注意， $f_X(x)$ 还是概率密度函数，而不是概率。也就是说， $f_{X,Y}(x,y)$ 二重积分得到概率， $f_{X,Y}(x,y)$ “偏积分”得到的还是概率密度函数。

图 12 比较 $f_{X,Y}(x,y = c)$ 和 $f_X(x)$ 曲线。当 $y = c$ 取不同值时，我们可以看到 $f_{X,Y}(x,y)$ 和 $f_X(x)$ 曲线形状不同。当 $y = c$ 时， $f_{X,Y}(x,y = c)$ 不是一元连续随机变量 PDF；原因就是面积不为 1。但是经过归一化之后，它们就变成了一元随机变量 PDF。这个归一化的工具就是“贝叶斯定理”。

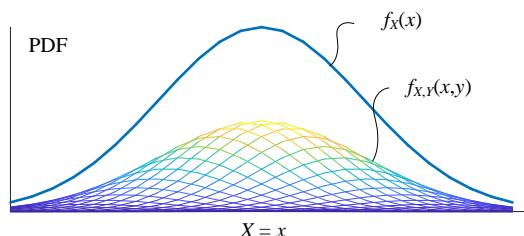


图 12. 比较联合概率密度 $f_{X,Y}(x,y)$ 和边缘概率密度 $f_X(x)$ 曲线

体密度 vs 面密度 vs 线密度

几何上来看，如图 13 所示， $f_{X,Y,Z}(x,y,z)$ 相当于“体密度”， $f_{X,Y}(x,y)$ 相当于“面密度”， $f_X(x)$ 相当于“线密度”。而概率值就相当于质量。

用白话说，体密度就是“铁块”的密度，计算铁块质量时会用到“体积 × 体密度”。

面密度就是“铁皮”的密度。铁皮厚度太薄，不便测量。计算铁皮质量时，我们用“面积 × 面密度”。

线密度对应“铁丝”的密度。关心铁丝横截面面积没有意义，实践中铁丝粗细有特定标准、型号。计算铁丝质量时，我们用“长度 × 线密度”。

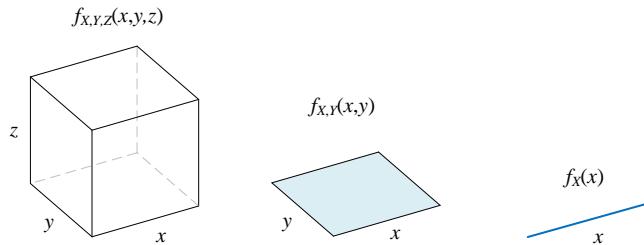


图 13. 体密度、面密度、线密度

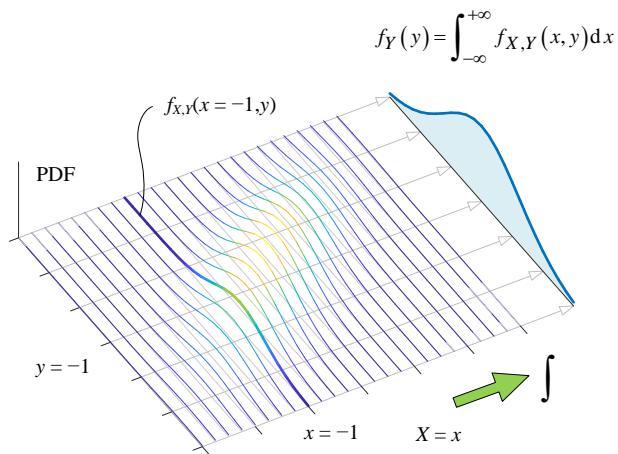
边缘概率密度函数 $f_Y(y)$

同理，如图 14 所示，连续随机变量 Y 的边缘分布概率密度函数 $f_Y(y)$ 可以通过 $f_{X,Y}(x,y)$ 对 x “偏积分”得到：

$$\underbrace{f_Y(y)}_{\text{Marginal}} = \overbrace{\int_{-\infty}^{+\infty} \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} dx}^{\text{Eliminate } x} \quad (25)$$

上式相当消去了变量 x 。上式也可以简写为：

$$\underbrace{f_Y(y)}_{\text{Marginal}} = \int_x^{\underbrace{\int_{-\infty}^{+\infty} f_{X,Y}(x,y) dx}_{\text{Joint}}} \underbrace{\text{Eliminate } x}_{(26)}$$

图 14. $f_{X,Y}(x,y)$ 对 x “偏积分”得到边缘分布概率密度函数 $f_Y(y)$

6.5 条件概率：引入贝叶斯定理

条件概率密度函数 $f_{X|Y}(x|y)$

设 X 和 Y 为连续随机变量，联合概率密度函数为 $f_{X,Y}(x,y)$ 。利用贝叶斯定理，在给定 $Y=y$ 条件下，且 $f_Y(y) > 0$ ， X 的条件概率密度函数 $f_{X|Y}(x|y)$ 为：

$$\underbrace{f_{X|Y}(x|y)}_{\text{Conditional}} = \frac{\overbrace{f_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{f_Y(y)}_{\text{Marginal}}} \quad (27)$$

⚠ 再次强调，上式中，边缘 $f_Y(y)$ 也是概率密度。

图 15 中 $f_{X,Y}(x,y = -1)$ 曲线代表 $Y = -1$ 时 (X, Y) 联合概率密度函数。

$f_{X,Y}(x,y = -1)$ 对 x 在 $(-\infty, +\infty)$ 积分的结果为边缘概率密度 $f_Y(y = -1)$ 。也就是说， $f_{X,Y}(x,y = -1)$ 曲线面积为边缘概率密度 $f_Y(y = -1)$ 。

下一步， $f_{X,Y}(x,y = -1)$ 经过 $f_Y(y = -1)$ 缩放得到条件概率曲线 $f_{X|Y}(x|y = -1)$ 。

⚠ 注意， $f_{X|Y}(x|y = -1)$ 和横轴围成图形的面积为 1，这代表 $Y = -1$ 这个新的样本空间概率为 1。

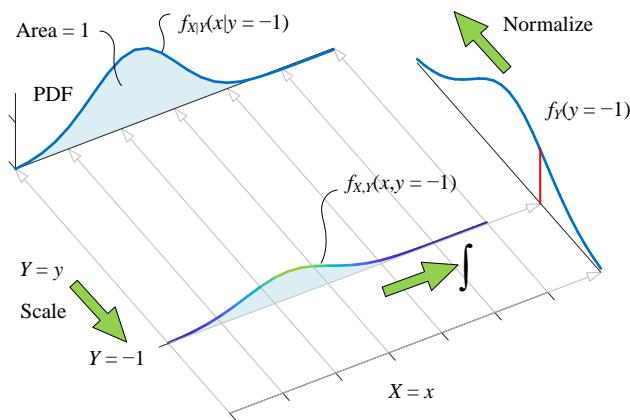
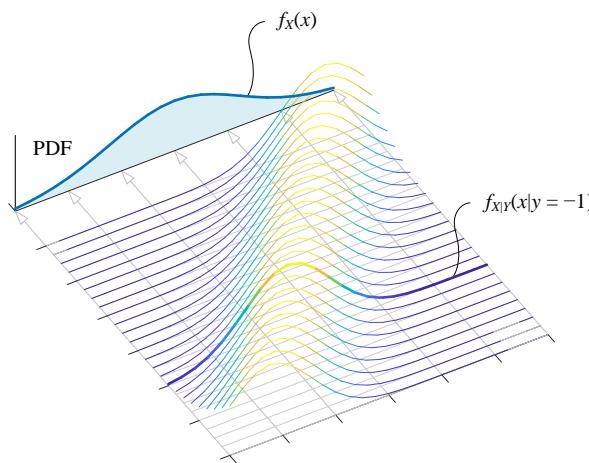
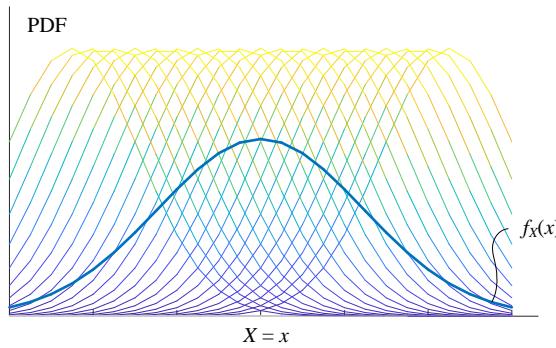
图 15. 给定 $Y=y$ 条件下且 $f_Y(y) > 0$, X 的条件概率密度函数

图 16 比较 $f_X(x)$ 和 y 取不同值时条件概率密度函数 $f_{X|Y}(x|y)$ 图像。将这些曲线投影到同一个平面，得到图 17。注意，图 17 中所有曲线和横轴围成图形的面积都是 1。

图 16. 比较边缘概率密度 $f_X(x)$ 和条件概率密度 $f_{X|Y}(x|y)$ 图 17. 比较边缘概率密度 $f_X(x)$ 和条件概率密度 $f_{X|Y}(x|y)$, 投影在平面上

条件概率密度函数 $f_{Y|X}(y|x)$

给定 $X = x$ 条件下，且 $f_X(x) > 0$ ，条件概率密度函数 $f_{Y|X}(y|x)$ 可以通过下式求得：

$$\underbrace{f_{Y|X}(y|x)}_{\text{Conditional}} = \frac{\overbrace{f_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{f_X(x)}_{\text{Marginal}}} \quad (28)$$

如图 18 所示为，当 $X = -1$ 条件下，联合概率密度函数 $f_{X,Y}(x = -1, y)$ 首先对 y 在 $(-\infty, +\infty)$ 积分的结果为边缘概率密度值 $f_X(x = -1)$ 。下一步， $f_{X,Y}(x = -1, y)$ 经过 $f_X(x = -1)$ 缩放得到条件概率曲线 $f_{Y|X}(y|x = -1)$ 。

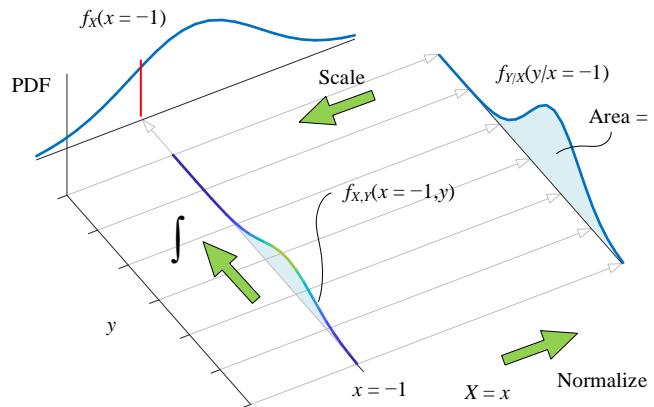


图 18. 给定 $X = x$ 条件下且 $f_X(x) > 0$ ， Y 的条件概率密度函数

图 19 比较 $f_Y(y)$ 和 x 取不同值时条件概率密度函数 $f_{Y|X}(y|x)$ 图像。

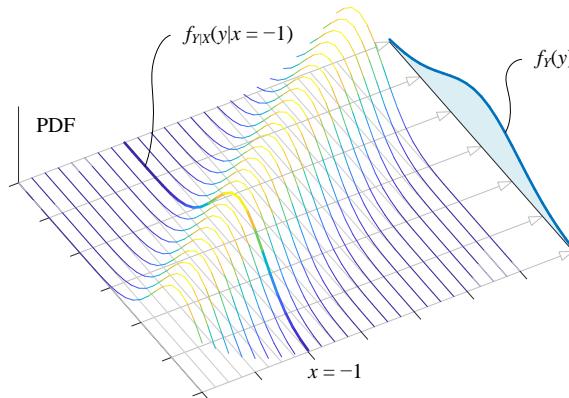


图 19. 比较边缘概率密度 $f_Y(y)$ 和条件概率密度 $f_{Y|X}(y|x)$ 图像

联合概率、边缘概率、条件概率

根据贝叶斯定理，联合概率、边缘概率、条件概率三者关系为：

$$\underbrace{f_{X,Y}(x,y)}_{\text{Joint}} = \underbrace{f_{X|Y}(x|y)}_{\text{Conditional}} \underbrace{f_Y(y)}_{\text{Marginal}} = \underbrace{f_{Y|X}(y|x)}_{\text{Conditional}} \underbrace{f_X(x)}_{\text{Marginal}} \quad (29)$$

在(23)基础上，连续随机变量 X 的边缘分布概率密度函数 $f_X(x)$ 可以通过下式获得：

$$\underbrace{f_X(x)}_{\text{Marginal}} = \int_{-\infty}^{+\infty} \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} dy = \int_{-\infty}^{+\infty} \underbrace{f_{X|Y}(x|t)}_{\text{Conditional}} f_Y(t) dt \quad (30)$$

同理，连续随机变量 Y 的边缘分布概率密度函数 $f_Y(y)$ 可以通过下式计算得到：

$$\underbrace{f_Y(y)}_{\text{Marginal}} = \int_{-\infty}^{+\infty} \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} dx = \int_{-\infty}^{+\infty} \underbrace{f_{Y|X}(y|s)}_{\text{Conditional}} f_X(s) ds \quad (31)$$

6.6 独立性：比较条件概率和边缘概率

如果连续随机变量 X 和 Y 独立，下式成立：

$$f_{X|Y}(x|y) = f_X(x) \quad (32)$$

图 20 所示为 X 和 Y 独立，条件概率密度函数 $f_{X|Y}(x|y)$ 和边缘概率密度函数 $f_X(x)$ 之间关系。我们发现条件概率 $f_{X|Y}(x|y)$ 的曲线和 Y 的取值无关。条件概率 $f_{X|Y}(x|y)$ 的曲线形状和边缘概率 $f_X(x)$ 完全一致。这和图 16 完全不同。

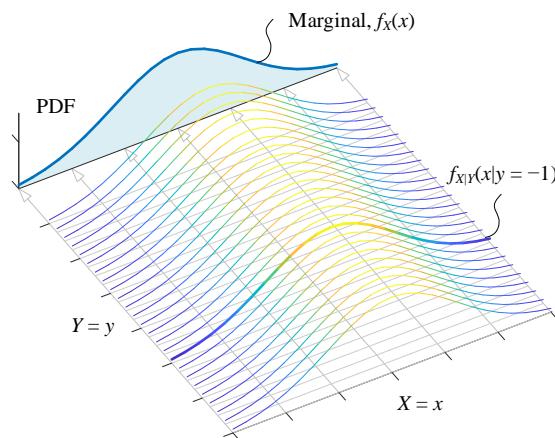


图 20. X 和 Y 独立，条件概率 $f_{X|Y}(x|y)$ 和边缘概率 $f_X(x)$ 之间关系

(32) 等价于：

$$f_{Y|X}(y|x) = f_Y(y) \quad (33)$$

图 21 所示为 X 和 Y 独立，条件概率 $f_{Y|X}(y|x)$ 和边缘概率 $f_Y(y)$ 的图像完全一致。

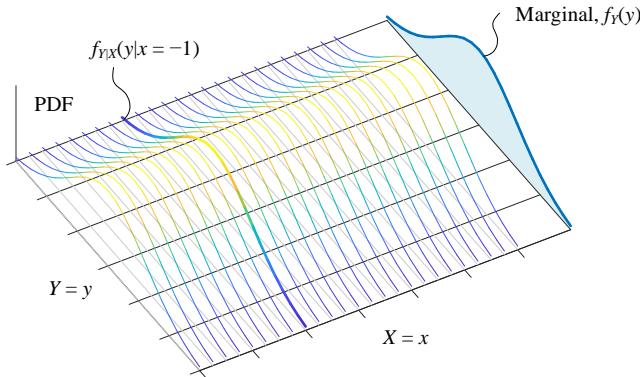


图 21. X 和 Y 独立，条件概率 $f_{Y|X}(y|x)$ 和边缘概率 $f_Y(y)$ 之间关系

独立：联合概率

对于两个连续随机变量 X 和 Y ，如果两者独立，则联合概率密度函数 $f_{X,Y}(x,y)$ 为边缘概率密度函数 $f_X(x)$ 和 $f_Y(y)$ 的乘积：

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad (34)$$

图 22 所示为连续随机变量 X 和 Y 独立，联合概率 $f_{X,Y}(x,y)$ 曲面。图 23 所示为联合概率 $f_{X,Y}(x,y)$ 平面等高线。

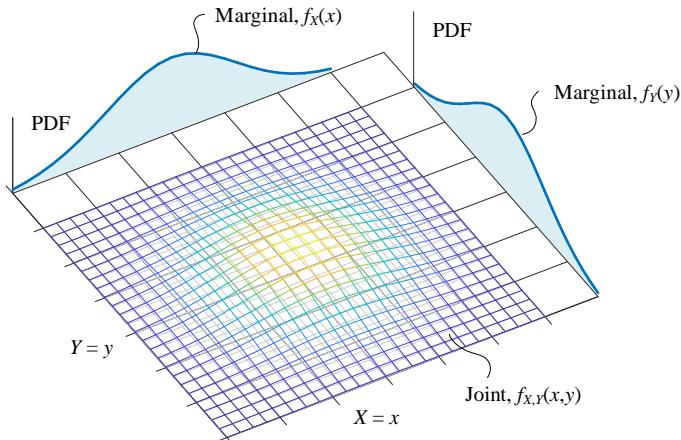
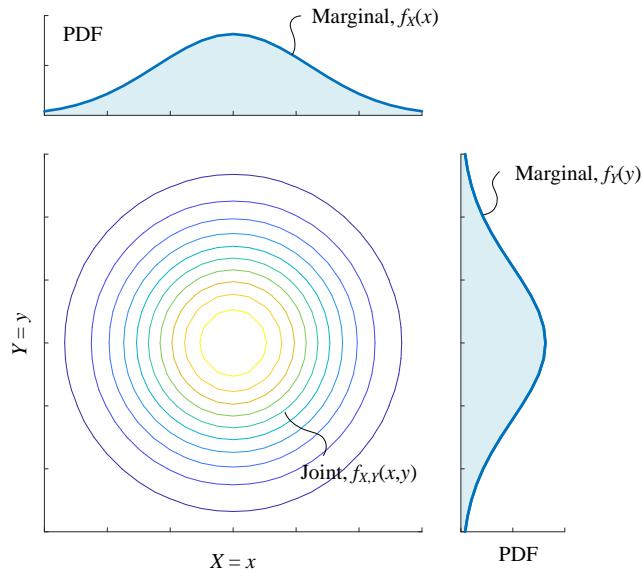


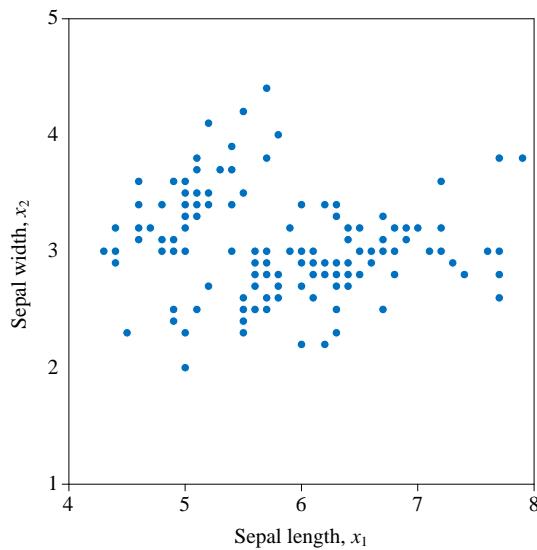
图 22. 连续随机变量 X 和 Y 独立，联合概率密度 $f_{X,Y}(x,y)$ 曲面

图 23. 连续随机变量 X 和 Y 独立，联合概率密度 $f_{X,Y}(x,y)$ 曲面等高线

6.7 以鸢尾花数据为例：不考虑分类标签

本章以下两节还是用鸢尾花数据集花萼长度 (X_1)、花萼宽度 (X_2)、分类标签 (Y) 为例，讲解本章前文介绍连续随机变量主要知识点。图 24 所示为不考虑分类时，鸢尾花样本数据花萼长度、花萼宽度散点图。

这两节采用和第 5 章 9、10 两节几乎一样的结构，方便大家对照阅读。



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 24. 鸢尾花数据花萼长度、花萼宽度散点图，不考虑分类

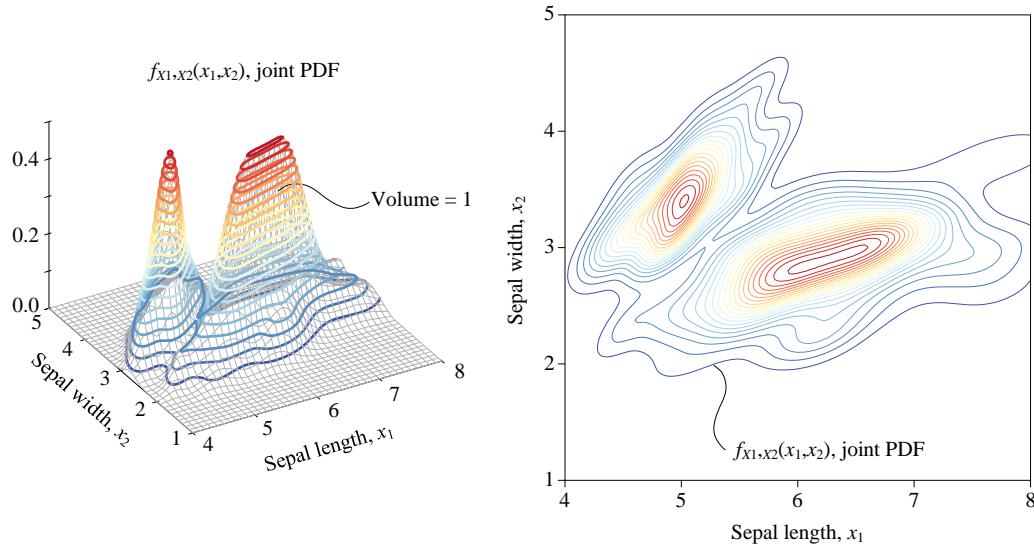
概率密度估计 → 联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$

基于高斯**核密度估计** (kernel density estimation, KDE)，我们可以得到如图 25 所示联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 。暖色系对应较大的概率密度值，也就是说鸢尾花样本分布更为密集。

核密度估计的基本思想是，通过在每个数据点处放置一个核函数（如高斯核函数），以此来估计概率密度函数。这样，在整个数据集上使用核函数后，我们可以获得一条连续的概率密度曲线，该曲线可以用来估计各种统计量，如均值和方差。



再次强调，图 25 仅仅代表 $f_{X_1, X_2}(x_1, x_2)$ 的一种估计。即便采用相同的 KDE，使用不同的核函数、改变算法参数会导致 $f_{X_1, X_2}(x_1, x_2)$ 曲面形状变化。本书第 18 章将专门讲解核密度估计方法。

图 25. 联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 三维等高线和平面等高线，不考虑分类

举个例子，花萼长度 (X_1) 为 6.5、花萼宽度 (X_2) 为 2.0 时，联合概率密度估计为：

$$\underbrace{f_{X_1, X_2}(x_1 = 6.5, x_2 = 2.0)}_{\text{Joint PDF}} \approx 0.02097 \quad (35)$$

⚠ 注意，0.02097 这个数值是概率密度，不是概率。也就是说，我们不能说鸢尾花取到花萼长度 (X_1) 为 6.5、花萼宽度 (X_2) 为 2.0 时对应的概率值为 0.02097，即便这个值某种程度上也代表可能性。

由于 $f_{X_1, X_2}(x_1, x_2)$ 有两个随机变量，对它二重积分可以得到概率值。二重积分就相当于“穷举法”。

采用“穷举法”，图 25 中 $f_{X_1, X_2}(x_1, x_2)$ 曲面和整个水平面围成的几何形体体积为 1，即：

$$\int \int_{x_2, x_1} f_{X_1, X_2}(x_1, x_2) d x_1 d x_2 = \frac{1}{\text{Probability}} \quad (36)$$

联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 的剖面线

$f_{X_1, X_2}(x_1, x_2)$ 本质上是个二元函数。



《数学要素》第 10 章介绍过除了等高线，我们还可以使用“剖面线”分析二元函数。

如图 26 所示，当固定 x_1 取值时， $f_{X_1, X_2}(x_1 = c, x_2)$ 代表一条曲线。将一系列类似曲线投影到竖直平面得到图 26 (b)。图 26 (b)，这些直线和整个水平轴围成的面积就是边缘概率 $f_{X_1}(x_1 = c)$ 。而计算面积的数学工具就是“偏积分”。

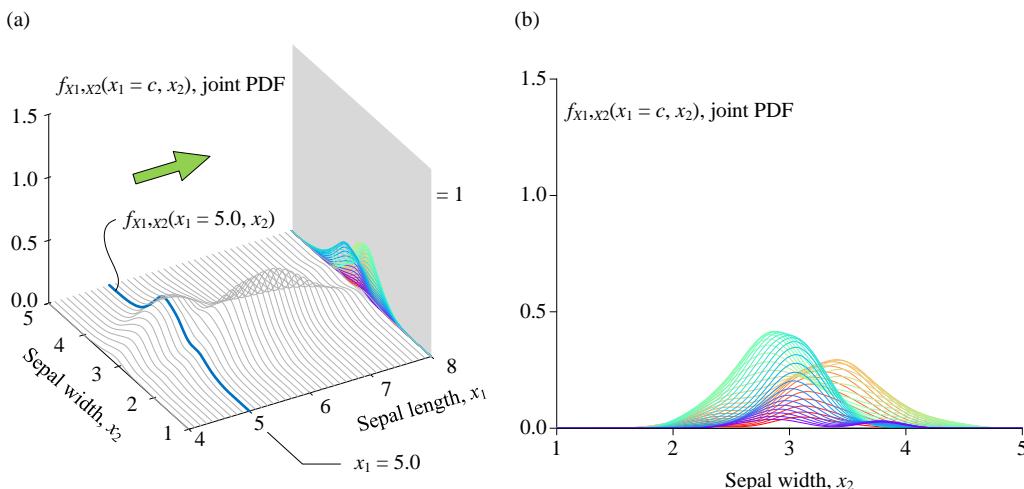
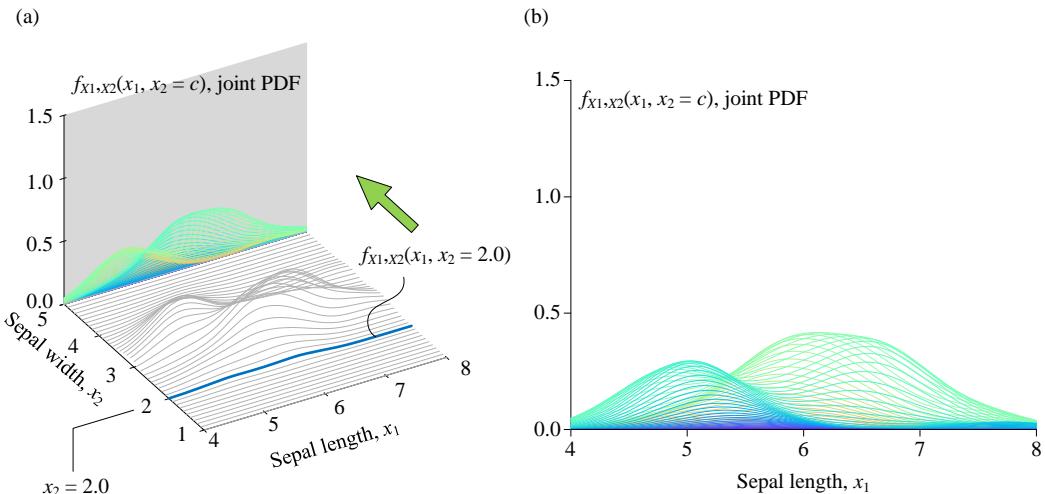


图 26. 固定 x_1 时，概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 随 x_2 变化

图 27 所示为固定 x_2 时，概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 随 x_1 变化。图 26 (b) 中直线和整个水平轴围成的面积对应边缘概率 $f_{X_2}(x_2 = c)$ 。

图 27. 固定 x_2 时，概率密度函数 $f_{X1,X2}(x_1, x_2)$ 随 x_1 变化

花萼长度边缘 PDF $f_{X1}(x_1)$: 偏积分

图 28 所示为求解花萼长度边缘概率密度函数 $f_{X1}(x_1)$ 的过程：

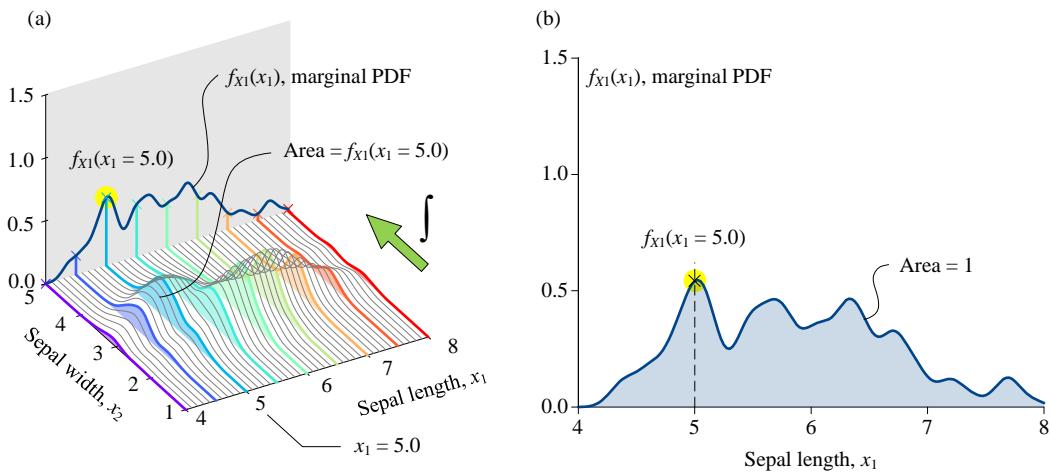
$$\underbrace{f_{X1}(x_1)}_{\text{Marginal}} = \int_{x_2} \underbrace{f_{X1,X2}(x_1, x_2)}_{\text{Joint}} dx_2 \quad (37)$$

举个例子，当花萼长度 (X_1) 取值为 5.0 时，对应的边缘概率 $f_{X1}(5.0)$ 可以通过如下偏积分得到：

$$f_{X1}(x_1 = 5.0) = \int_{x_2} f_{X1,X2}(x_1 = 5.0, x_2) dx_2 \quad (38)$$

图 28 中彩色阴影面积对应边缘概率，即 $f_{X1}(x_1)$ 曲线特定一点的高度。再次强调， $f_{X1}(x_1)$ 本身也是概率密度，不是概率值。 $f_{X1}(x_1)$ 再积分可以得到概率。

如图 28 (b) 所示， $f_{X1}(x_1)$ 曲线和整个横轴围成图形的面积为 1。大家可以试着用数值积分计算期望值 $E(X_1)$ 。

图 28. 偏积分求解边缘概率 $f_{X1}(x_1)$

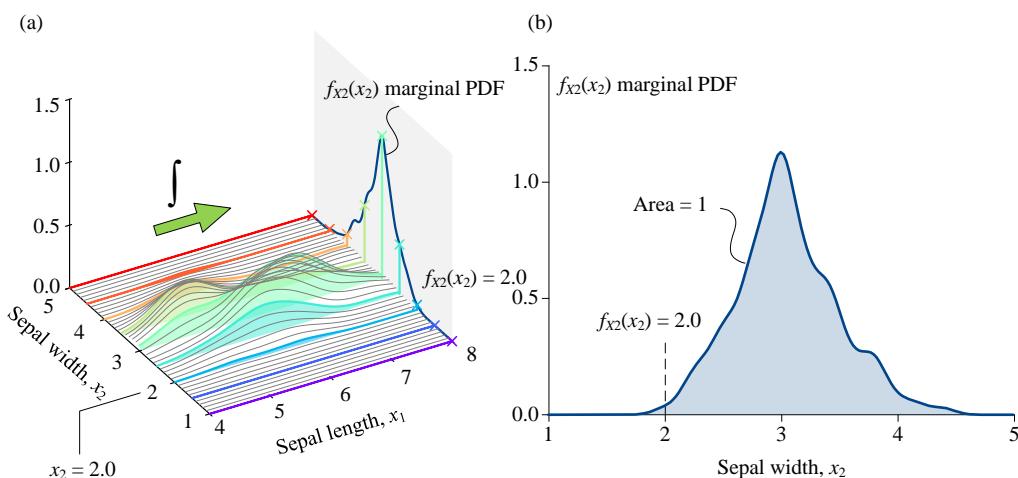
花萼宽度边缘 PDF $f_{X2}(x_2)$: 偏求和

图 29 所示为求解花萼宽度边缘概率密度函数的过程：

$$\underbrace{f_{X2}(x_2)}_{\text{Marginal}} = \int_{x_1} \underbrace{f_{X1,X2}(x_1, x_2)}_{\text{Joint}} dx_1 \quad (39)$$

举个例子，当花萼宽度 (X_2) 取值为 2.0 时，对应的边缘概率密度 $f_{X2}(2.0)$ 可以通过如下偏积分得到：

$$f_{X2}(x_2 = 2.0) = \int_{x_1} f_{X1,X2}(x_1, x_2 = 2.0) dx_1 \quad (40)$$

图 29. 偏积分求解边缘概率 $f_{X2}(x_2)$

联合 PDF vs 边缘 PDF

图 30 所示为联合 PDF 和边缘 PDF 之间关系。图中联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 采用高斯 KDE 估计得到。图 30 中的 $f_{X_1, X_2}(x_1, x_2)$ 比较精准地捕捉到了鸢尾花样本数据的分布特征。

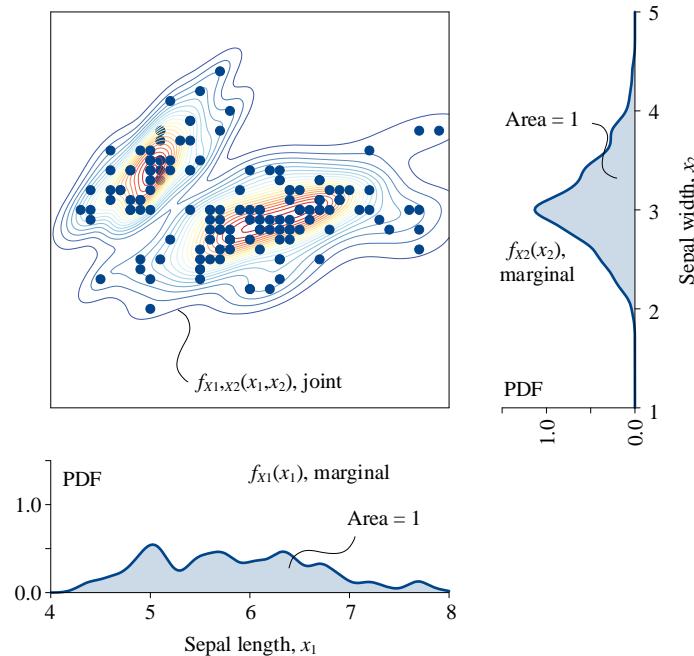


图 30. 联合 PDF 和边缘 PDF 之间关系

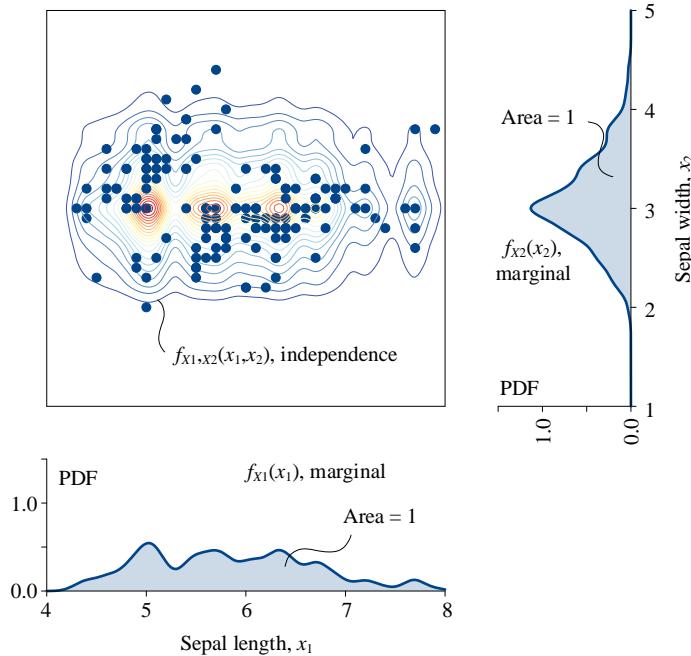
假设独立

如果假设 X_1 和 X_2 独立，联合概率密度 $f_{X_1, X_2}(x_1, x_2)$ 可通过下式计算得到：

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \quad (41)$$

图 31 所示为假设 X_1 和 X_2 独立时 $f_{X_1, X_2}(x_1, x_2)$ 的平面等高线和边缘 PDF 之间关系。

比较鸢尾花样本数据分布和假设 X_1 和 X_2 独立时估算得到的 $f_{X_1, X_2}(x_1, x_2)$ 等高线，很遗憾地发现图 31 这个联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 没有合理反映样本数据分布，尽管图 30 和图 31 边缘概率完全一致。

图 31. 联合概率，假设 X_1 和 X_2 独立

给定花萼长度，花萼宽度的条件 PDF $f_{X_2|X_1}(x_2|x_1)$

如图 32 所示，利用贝叶斯定理，条件概率密度 $f_{X_2|X_1}(x_2|x_1)$ 可以通过下式计算：

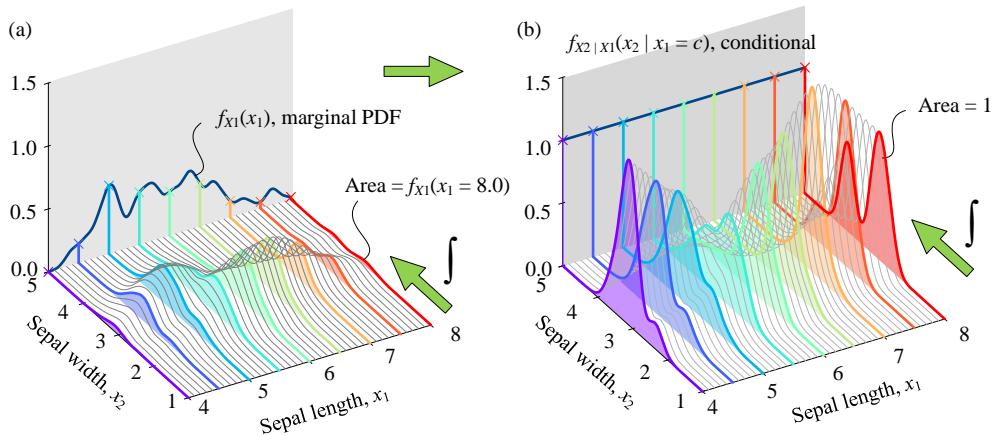
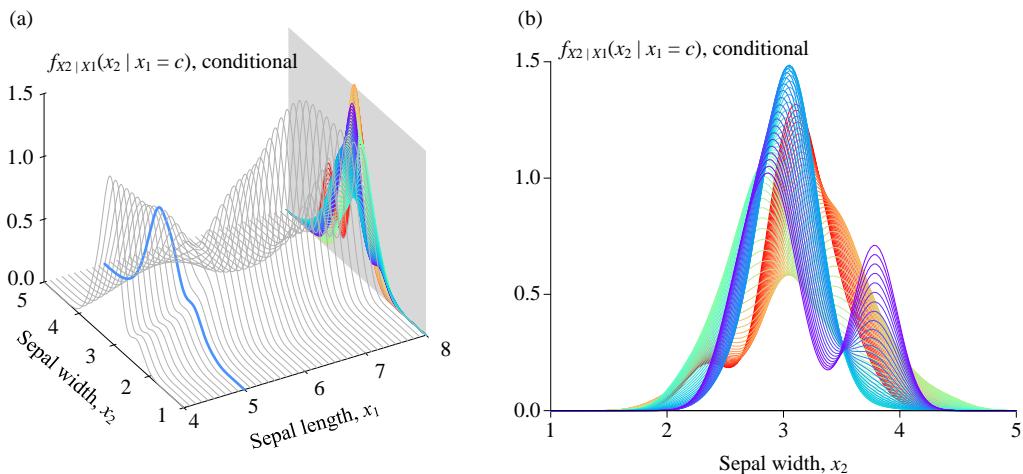
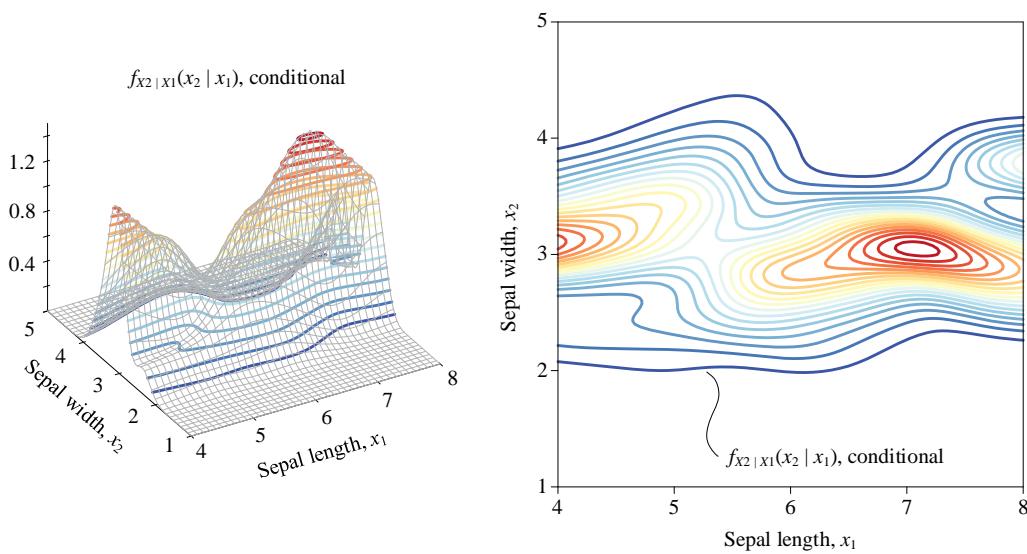
$$\underbrace{f_{X_2|X_1}(x_2|x_1)}_{\text{Conditional}} = \frac{\overbrace{f_{X_1, X_2}(x_1, x_2)}^{\text{Joint}}}{\underbrace{f_{X_1}(x_1)}_{\text{Marginal}}} \quad (42)$$

⚠ 注意，上式中 $f_{X_1}(x_1) > 0$ 。上式分母中的边缘概率 $f_{X_1}(x_1)$ 起到归一化作用。

如图 32 (b) 所示，经过归一化的条件概率曲线围成的面积变为 1，即：

$$\int_{x_2} \underbrace{f_{X_2|X_1}(x_2|x_1)}_{\text{Conditional}} dx_2 = \int_{x_2} \frac{\overbrace{f_{X_1, X_2}(x_1, x_2)}^{\text{Joint}}}{\underbrace{f_{X_1}(x_1)}_{\text{Marginal}}} dx_2 = \frac{\int_{x_2} f_{X_1, X_2}(x_1, x_2) dx_2}{f_{X_1}(x_1)} = \frac{f_{X_1}(x_1)}{f_{X_1}(x_1)} = 1 \quad (43)$$

将不同位置的条件 PDF $f_{X_2|X_1}(x_2|x_1)$ 曲线投影到平面得到图 33。图 33 (b) 中每条曲线和横轴围成面积都是 1。请大家仔细比较图 26 和图 33。此外， $f_{X_2|X_1}(x_2|x_1)$ 本身也是一个二元函数。图 34 所示为 $f_{X_2|X_1}(x_2|x_1)$ 三维等高线和平面等高线。

图 32. 计算条件概率 $f_{X2|X1}(x_2 | x_1)$ 原理图 33. $f_{X2|X1}(x_2 | x_1)$ 曲线投影到平面图 34. $f_{X2|X1}(x_2 | x_1)$ 条件概率密度三维等高线和平面等高线，不考虑分类

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

给定花萼宽度，花萼长度的条件概率密度函数 $f_{X_1|X_2}(x_1 | x_2)$

如图 35 所示，同样利用贝叶斯定理，条件 PDF $f_{X_1|X_2}(x_1 | x_2)$ 可以通过下式计算：

$$\underbrace{f_{X_1|X_2}(x_1 | x_2)}_{\text{Conditional}} = \frac{\overbrace{f_{X_1, X_2}(x_1, x_2)}^{\text{Joint}}}{\underbrace{f_{X_2}(x_2)}_{\text{Marginal}}} \quad (44)$$

注意，上式中 $f_{X_2}(x_2) > 0$ 。

类似前文，(44) 中分母中 $f_{X_2}(x_2)$ 同样起到归一化作用。如图 35 (b) 所示，经过归一化 $f_{X_1|X_2}(x_1 | x_2)$ 面积变为 1，即：

$$\int_{x_1} \underbrace{f_{X_1|X_2}(x_1 | x_2)}_{\text{Conditional}} dx_1 = \int_{x_1} \frac{\overbrace{f_{X_1, X_2}(x_1, x_2)}^{\text{Joint}}}{\underbrace{f_{X_2}(x_2)}_{\text{Marginal}}} dx_1 = \frac{\int_{x_1} f_{X_1, X_2}(x_1, x_2) dx_1}{f_{X_2}(x_2)} = \frac{f_{X_2}(x_2)}{f_{X_2}(x_2)} = 1 \quad (45)$$

将不同位置的条件概率密度 $f_{X_1|X_2}(x_1 | x_2)$ 曲线投影到平面得到图 36。图 36 (b) 中每条曲线和横轴围成面积都是 1。也请大家仔细比较图 27 和图 36。

$f_{X_1|X_2}(x_1 | x_2)$ 同样也是一个二元函数，如图 37 所示的 $f_{X_1|X_2}(x_1 | x_2)$ 三维等高线和平面等高线。

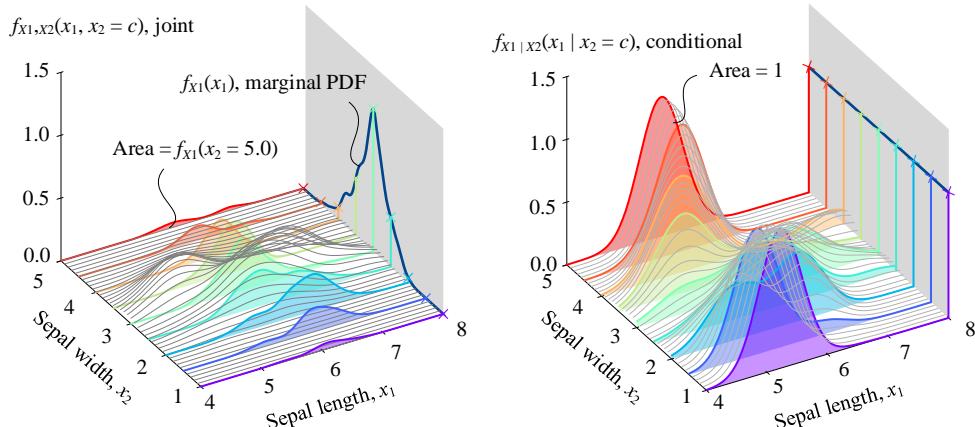
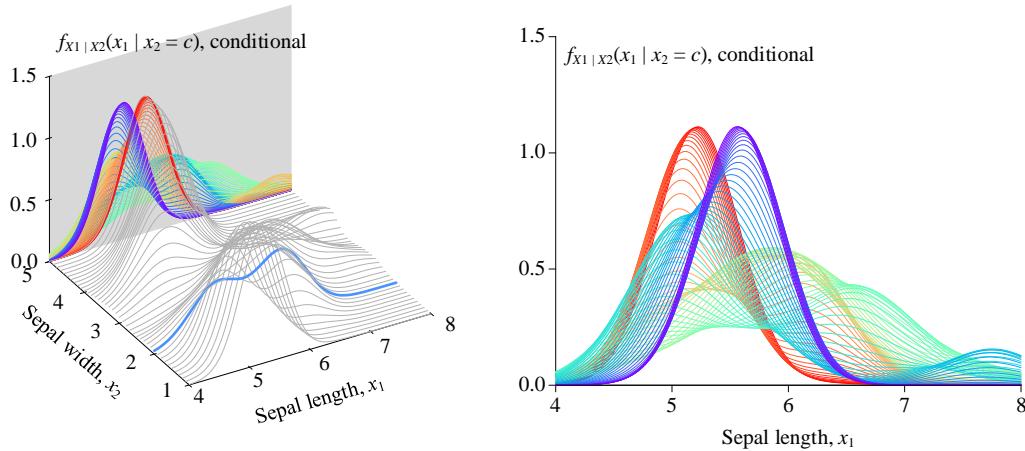
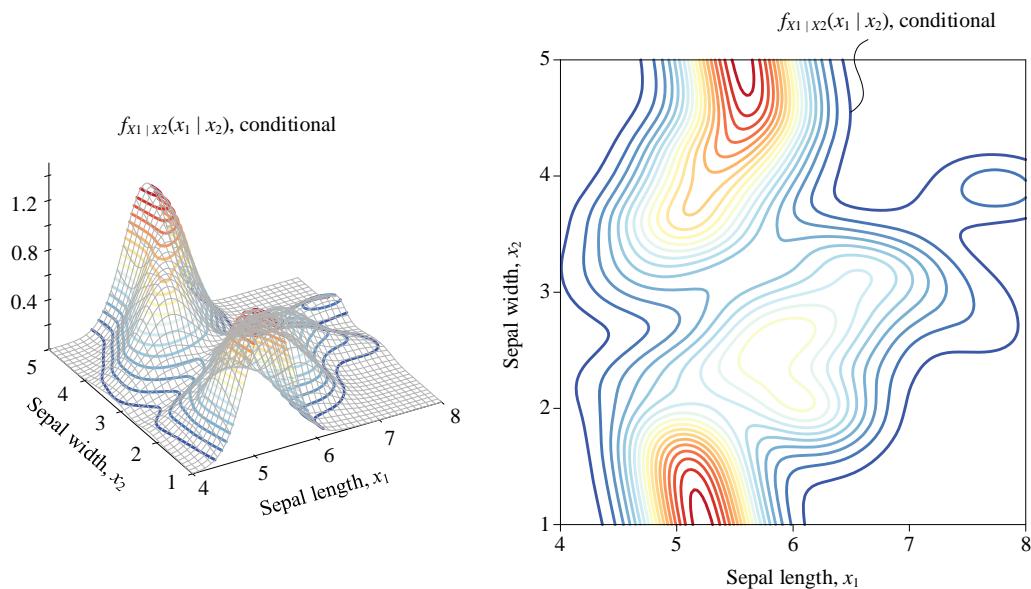


图 35. 计算条件概率 $f_{X_1|X_2}(x_1 | x_2)$ 原理

图 36. $f_{X1|X2}(x_1 | x_2)$ 曲线投影到平面图 37. $f_{X1|X2}(x_1 | x_2)$ 条件概率密度三维等高线和平面等高线，不考虑分类

6.8 以鸢尾花数据为例：考虑分类标签

本节将以鸢尾花标签为条件讨论条件概率。图 38 所示为考虑分类标签的鸢尾花数据散点图。

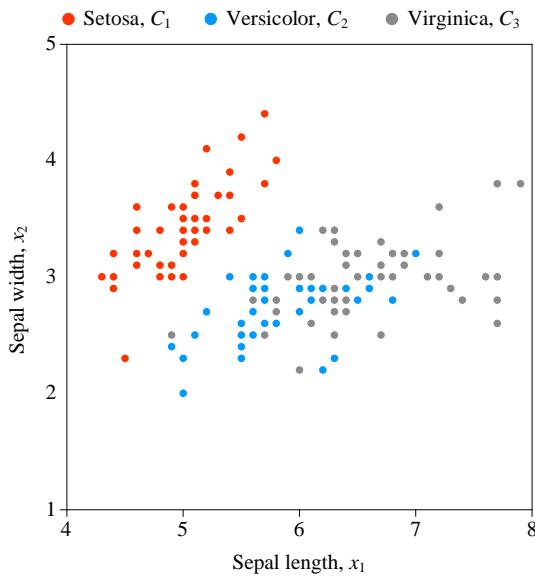


图 38. 鸢尾花数据萼片长度、萼片宽度散点图，考虑分类

给定分类标签 $Y = C_1$ (setosa)

图 39 所示为给定分类标签 $Y = C_1$ (setosa) 条件下，条件概率 $f_{X_1, X_2 | Y}(x_1, x_2 | y = C_1)$ 平面等高线和条件边缘概率密度曲线。

$f_{X_1, X_2 | Y}(x_1, x_2 | y = C_1)$ 曲面和整个水平面围成体积为 1，也就是说：

$$\int \int \underbrace{f_{X_1, X_2 | Y}(x_1, x_2 | C_1)}_{\text{Conditional PDF}} d x_1 d x_2 = \frac{1}{\text{Probability}} \quad (46)$$

用 KDE 估算 $f_{X_1, X_2 | Y}(x_1, x_2 | y = C_1)$ 时，我们仅仅考虑标签为 C_1 的数据。同理，估算条件边缘概率曲线 $f_{X_1 | Y}(x_1 | y = C_1)$ 、 $f_{X_2 | Y}(x_2 | y = C_1)$ 时，我们也不考虑其他标签数据。

图 39 中， $f_{X_1 | Y}(x_1 | y = C_1)$ 、 $f_{X_2 | Y}(x_2 | y = C_1)$ 分别和 x_1 、 x_2 围成的面积也是 1，即：

$$\begin{aligned} \int \underbrace{f_{X_1 | Y}(x_1 | C_1)}_{\text{Conditional PDF}} d x_1 &= \frac{1}{\text{Probability}} \\ \int \underbrace{f_{X_2 | Y}(x_2 | C_1)}_{\text{Conditional PDF}} d x_2 &= \frac{1}{\text{Probability}} \end{aligned} \quad (47)$$

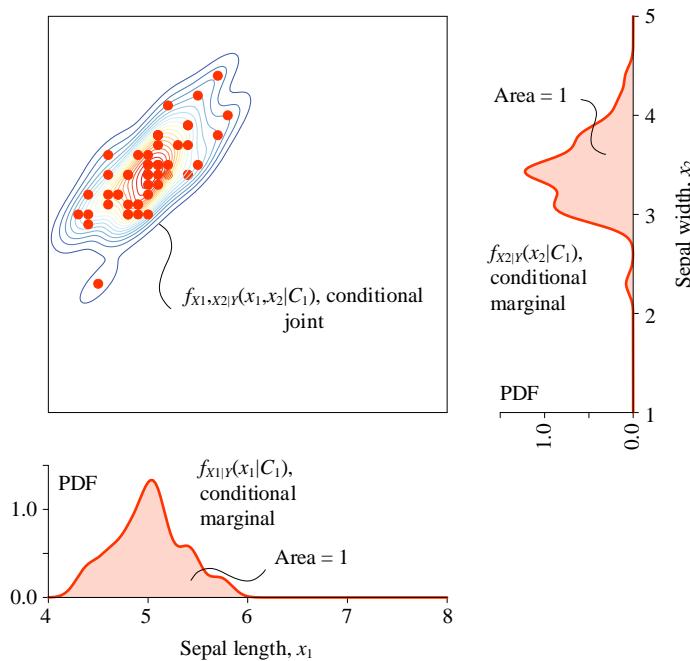


图 39. 条件概率 $f_{X1,X2|Y}(x_1, x_2 | y = C_1)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_1$ (setosa)

给定分类标签 $Y = C_2$ (versicolor)

图 40 所示为，给定分类标签 $Y = C_2$ (versicolor)，条件概率 $f_{X1,X2|Y}(x_1, x_2 | y = C_2)$ 平面等高线和条件边缘概率密度曲线。请大家自行分析这幅图。

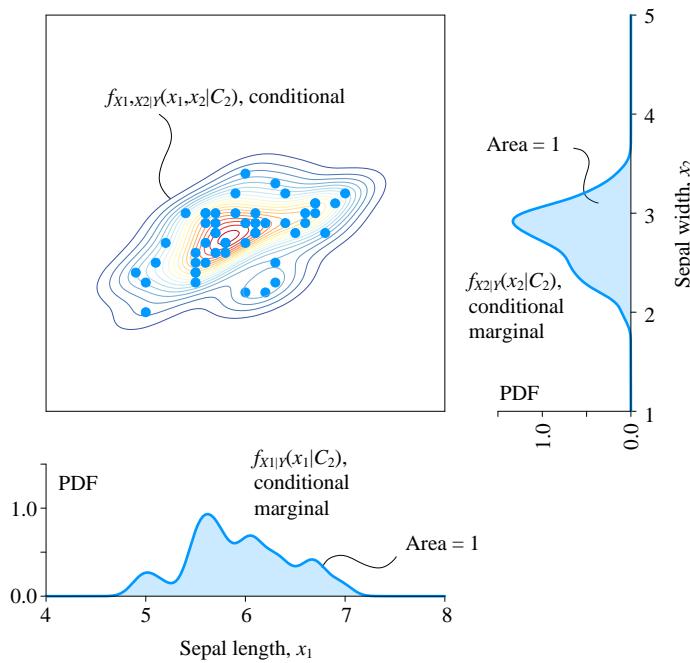


图 40. 条件 PDF $f_{X1,X2|Y}(x_1, x_2 | y = C_2)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_2$ (versicolor)

给定分类标签 $Y = C_3$ (virginica)

图 41 所示为，给定分类标签 $Y = C_3$ (virginica)，条件概率 $f_{X_1, X_2 | Y}(x_1, x_2 | y = C_3)$ 平面等高线和条件边缘概率密度曲线。也请大家自行分析这幅图。

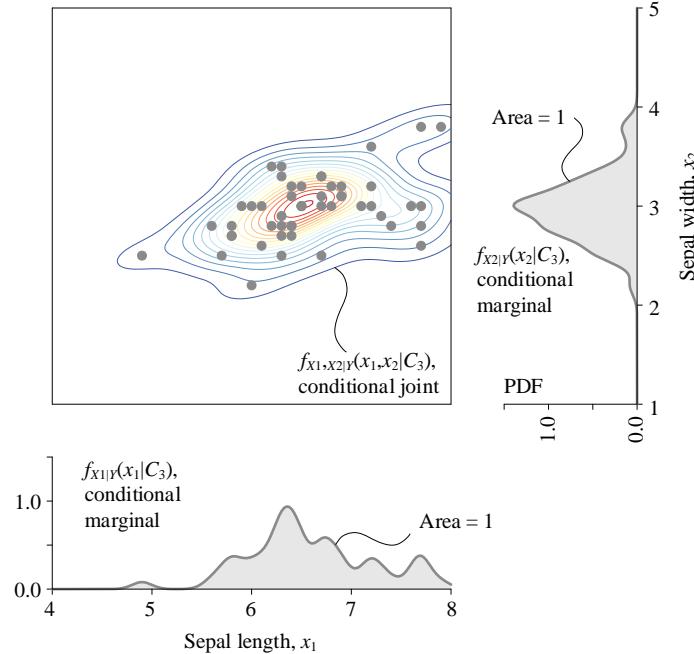


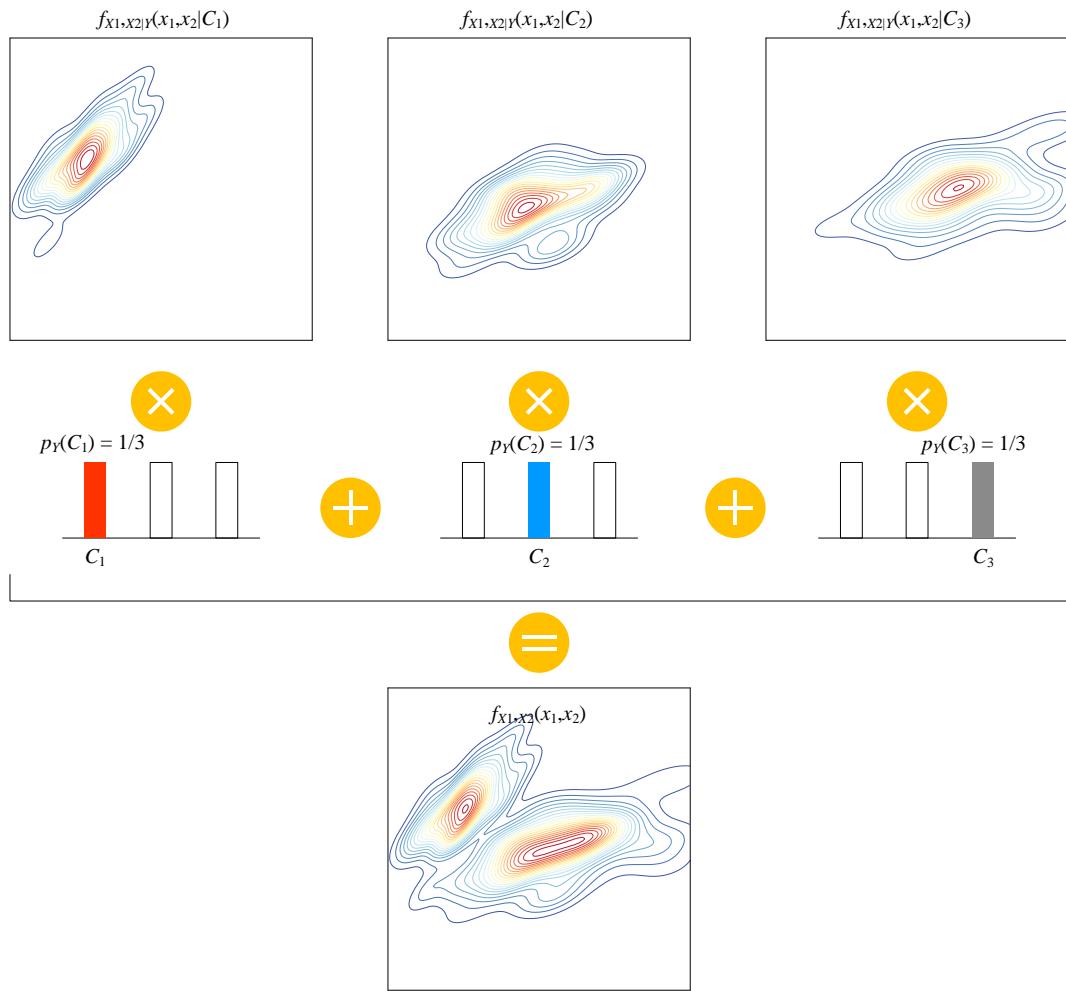
图 41. 条件 PDF $f_{X_1, X_2 | Y}(x_1, x_2 | y = C_3)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_3$ (virginica)

全概率定理：穷举法

如图 42 所示，利用全概率定理，三幅条件概率等高线叠加可以得到联合概率密度，即：

$$\begin{aligned} f_{X_1, X_2}(x_1, x_2) &= f_{X_1, X_2 | Y}(x_1, x_2 | y = C_1) p_Y(C_1) + \\ &\quad f_{X_1, X_2 | Y}(x_1, x_2 | y = C_2) p_Y(C_2) + \\ &\quad f_{X_1, X_2 | Y}(x_1, x_2 | y = C_3) p_Y(C_3) \end{aligned} \quad (48)$$

此外，请大家思考 $f_{X_1}(x_1)$ 、 $f_{X_1 | Y}(x_1 | y = C_1)$ 、 $f_{X_1 | Y}(x_1 | y = C_2)$ 、 $f_{X_1 | Y}(x_1 | y = C_3)$ 四者关系。

图 42. 利用全概率定理，计算 $f_{X1,X2}(x_1, x_2)$

给定 X_1 和 X_2 , Y 的条件概率：后验概率

根据贝叶斯定理，当 $f_{X1,X2}(x_1, x_2) > 0$ 时，**后验** (posterior) PDF $f_{Y|X1,X2}(C_k | x_1, x_2)$ 可以根据下式计算得到：

$$\overbrace{f_{Y|X1,X2}(C_k | x_1, x_2)}^{\text{Posterior}} = \frac{\overbrace{f_{X1,X2,Y}(x_1, x_2, C_k)}^{\text{Joint}}}{\underbrace{f_{X1,X2}(x_1, x_2)}_{\text{Evidence}}} \quad (49)$$

从分类角度来看，这相当于已知某个样本鸢尾花花萼长度和花萼宽度，该样本对应不同分类的概率。请大家修改代码自行绘制不同的后验概率 PDF 曲面。

→ 本书第 19、20 章将从这个角度探讨若何判定鸢尾花分类。

假设条件独立

如图 43 所示，如果假设条件独立， $f_{X_1, X_2|Y}(x_1, x_2 | y = C_1)$ 可以通过下式计算得到：

$$\underbrace{f_{X_1, X_2|Y}(x_1, x_2 | y = C_1)}_{\text{Conditional joint}} = \underbrace{f_{X_1|Y}(x_1 | y = C_1)}_{\text{Conditional marginal}} \cdot \underbrace{f_{X_2|Y}(x_2 | y = C_1)}_{\text{Conditional marginal}} \quad (50)$$

同理我们可以计算得到 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_2)$ 、 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_3)$ ，具体如图 44、图 45 所示。

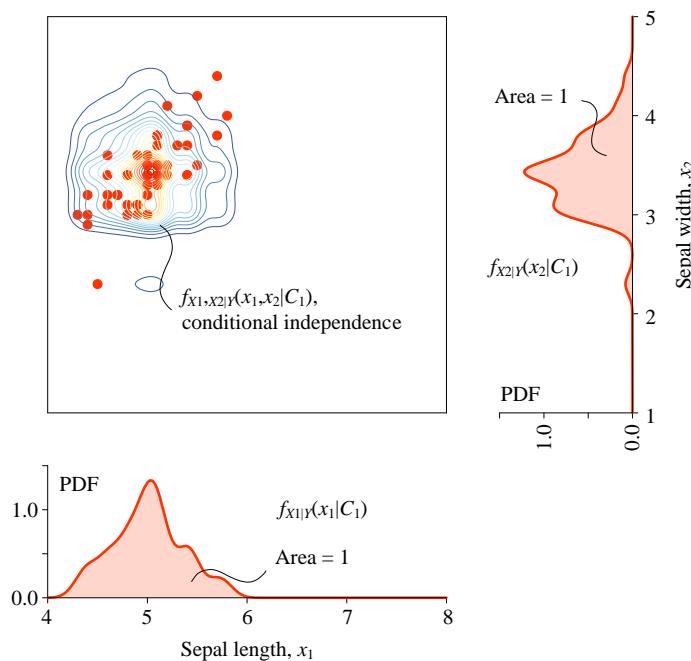
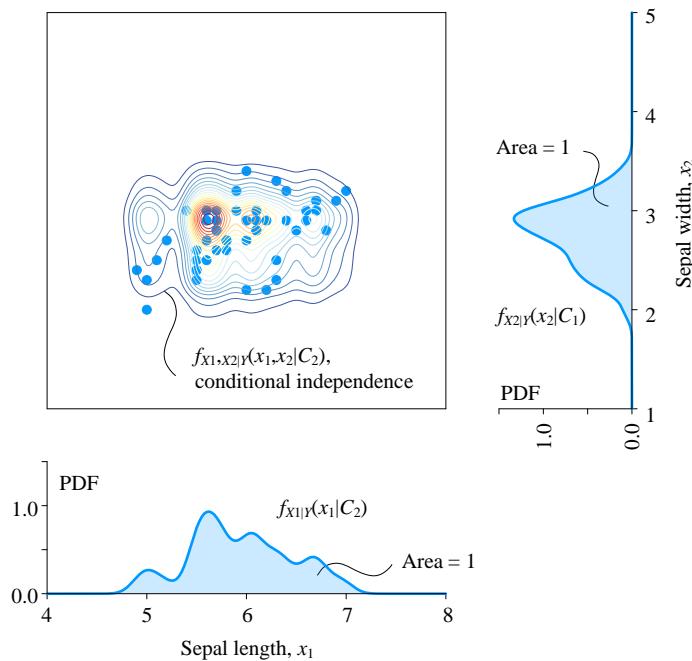
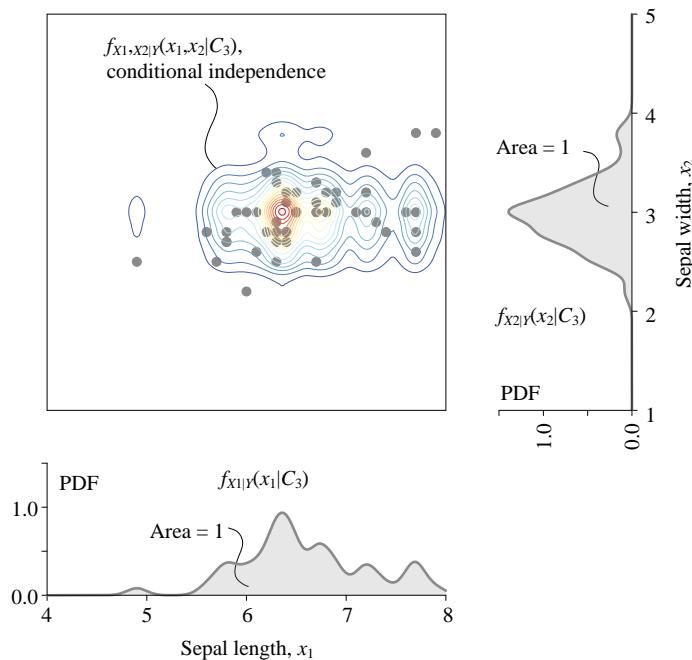


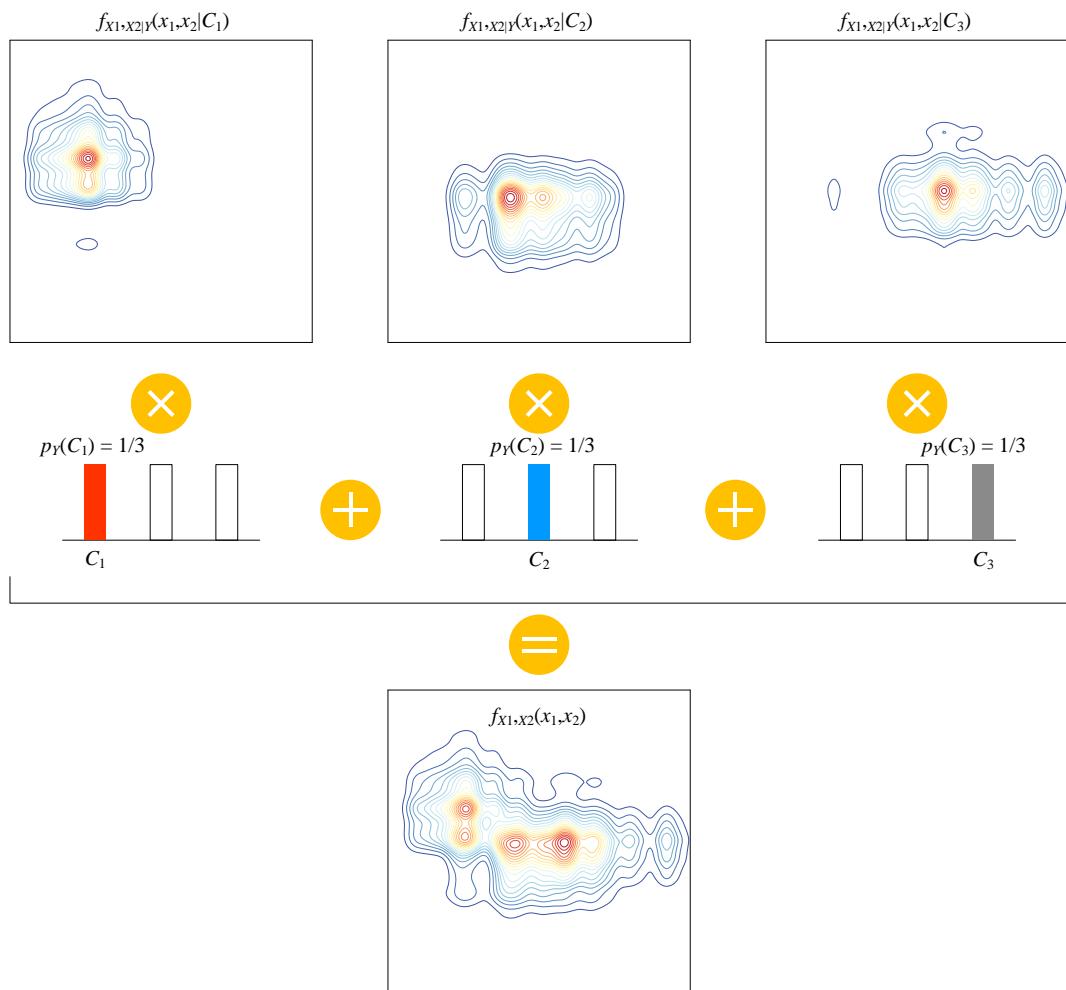
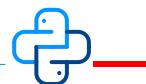
图 43. 给定 $Y = C_1$, X_1 和 X_2 条件独立, 估算条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_1)$

图 44. 给定 $Y = C_2$, X_1 和 X_2 条件独立, 估算条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_2)$ 图 45. 给定 $Y = C_3$, X_1 和 X_2 条件独立, 估算条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_3)$

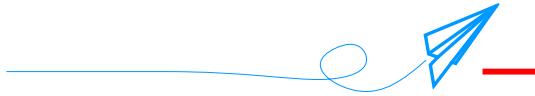
如图 46 所示, 并利用全概率定理, 我们也可以估算 $f_{X_1, X_2}(x_1, x_2)$:

$$\begin{aligned}
 f_{X_1, X_2}(x_1, x_2) &= f_{X_1, X_2|Y}(x_1, x_2 | y = C_1) p_Y(C_1) + \\
 &\quad f_{X_1, X_2|Y}(x_1, x_2 | y = C_2) p_Y(C_2) + \\
 &\quad f_{X_1, X_2|Y}(x_1, x_2 | y = C_3) p_Y(C_3) \\
 &= f_{X_1|Y}(x_1 | y = C_1) f_{X_2|Y}(x_2 | y = C_1) p_Y(C_1) + \\
 &\quad f_{X_1|Y}(x_1 | y = C_2) f_{X_2|Y}(x_2 | y = C_2) p_Y(C_2) + \\
 &\quad f_{X_1|Y}(x_1 | y = C_3) f_{X_2|Y}(x_2 | y = C_3) p_Y(C_3)
 \end{aligned} \tag{51}$$

→ 这是朴素贝叶斯分类器 (Naive Bayes classifier) 的重要技术细节之一。鸢尾花书《机器学习》一册将讲解朴素贝叶斯分类器。

图 46. 利用全概率定理，估算 $f_{X_1, X_2}(x_1, x_2)$ ，假设条件独立

Bk5_Ch06_01.py 绘制本章大部分图像。



为了帮助大家更容易发现离散随机变量、连续随机变量的区别和联系，本章最后特地做了如下表格。请大家逐行对比学习。下一章介绍常见连续随机变量的概率分布。

表 1. 比较离散和连续随机变量

	离散	连续
随机变量	取值可以一一列举出来，有限个或可数无穷个，比如 {0, 1}, {非负整数}	取值不可以一一列举出来，比如闭区间 [0, 1] 或 {非负实数}
一元随机变量概率质量/密度函数	概率质量函数 PMF, $p_X(x)$ PMF本身就是概率值 $0 \leq p_X(x) \leq 1$ 计算工具: Σ	概率密度函数 PDF, $f_X(x)$ PDF本身为概率密度 $0 \leq f_X(x)$ 注意 $f_X(x)$ 可以大于 1 计算工具: \int
归一化	$\sum_x p_X(x) = 1$	$\int_x f_X(x) dx = 1$
概率质量/密度函数图像	火柴梗图	曲线
计算概率 CDF	求和 $F_X(x) = \Pr(X \leq x) = \sum_{t \leq x} p_X(t)$	积分 $F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f_X(t) dt$
期望	$E(X) = \sum_x x \cdot p_X(x)$	$E(X) = \int_x x \cdot f_X(x) dx$
方差	$\text{var}(X) = \sum_x (x - E(X))^2 p_X(x)$	$\text{var}(X) = \int_x (x - E(X))^2 \cdot f_X(x) dx$
常见分布	离散均匀分布，伯努利分布，二项分布，多项分布，泊松分布，几何分布，超几何分布	连续均匀分布，高斯分布，逻辑分布，学生 t 分布，对数正态分布，指数分布，卡方分布，Beta 分布
二元随机变量联合概率	概率质量函数 PMF, $p_{X,Y}(x,y)$	概率密度函数 PDF, $f_{X,Y}(x,y)$
归一化	$\sum_{x_1} \sum_{x_2} p_{X_1, X_2}(x_1, x_2) = 1$	$\int_{x_2} \int_{x_1} f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 = 1$
边缘概率 求和法则	$p_{X,Y}(x,y)$ 偏求和结果为边缘 PMF $p_X(x) = \sum_y p_{X,Y}(x,y)$ $p_Y(y) = \sum_x p_{X,Y}(x,y)$	$f_{X,Y}(x,y)$ 偏积分结果为边缘 PDF $f_X(x) = \int_y f_{X,Y}(x,y) dy$ $f_Y(y) = \int_x f_{X,Y}(x,y) dx$
条件概率 $p_Y(y) > 0, p_X(x) > 0$ $f_Y(y) > 0, f_X(x) > 0$	$p_{X Y}(x y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$ $p_{Y X}(y x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$	$f_{Y X}(y x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ $f_{X Y}(x y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

条件概率归一化	$\sum_x p_{x y}(x y) = 1$ $\sum_y p_{y x}(y x) = 1$	$\int_x f_{x y}(x y) dx = 1$ $\int_y f_{y x}(y x) dy = 1$
随机变量独立	$p_{x y}(x y) = p_x(x)$ $p_{y x}(y x) = p_y(y)$	$f_{x y}(x y) = f_x(x)$ $f_{y x}(y x) = f_y(y)$
随机变量独立条件下，联合概率	$p_{x,y}(x,y) = p_x(x)p_y(y)$	$f_{x,y}(x,y) = f_x(x)f_y(y)$
随机变量条件独立，条件联合概率	$p_{x_1,x_2 y}(x_1,x_2 y) = p_{x_1 y}(x_1 y) \cdot p_{x_2 y}(x_2 y)$	$f_{x_1,x_2 y}(x_1,x_2 y) = f_{x_1 y}(x_1 y) \cdot f_{x_2 y}(x_2 y)$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

7 连续分布

分布相当于理想化假设



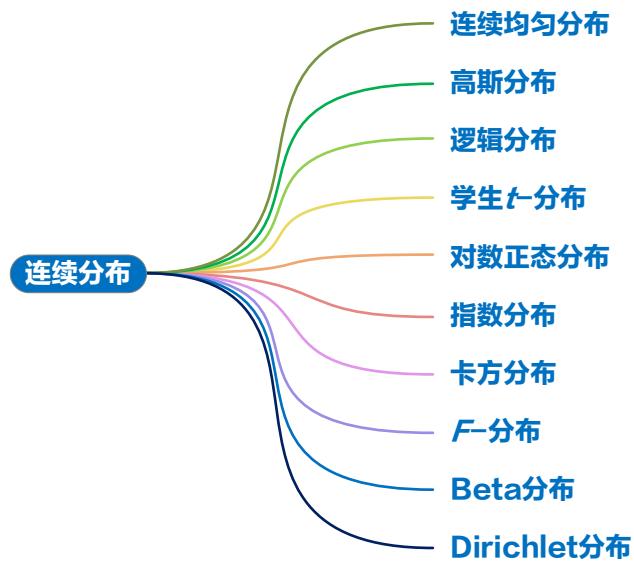
我们仅仅是，川流不息河水里的，一个个涡漩。肉体灰飞烟灭，潮流浩浩荡荡。

We are but whirlpools in a river of ever-flowing water. We are not the stuff that abides, but patterns that perpetuate themselves.

—— 诺伯特·维纳 (Norbert Wiener) | 美国数学家 | 1894 ~ 1964



- ◀ numpy.random.laplace() 拉普拉斯分布随机数发生器
- ◀ numpy.random.uniform() 均匀分布随机数发生器
- ◀ scipy.stats.beta() Beta 分布
- ◀ scipy.stats.beta.pdf() Beta 分布概率密度函数
- ◀ scipy.stats.chi2() 卡方分布函数
- ◀ scipy.stats.dirichlet() Dirichlet 分布
- ◀ scipy.stats.dirichlet.pdf() Dirichlet 分布概率密度函数
- ◀ scipy.stats.expon() 指数分布函数
- ◀ scipy.stats.laplace() 拉普拉斯分布函数
- ◀ scipy.stats.logistic() 逻辑分布函数
- ◀ scipy.stats.lognorm() 对数正态分布函数
- ◀ scipy.stats.norm() 正态分布函数
- ◀ scipy.stats.t() 学生 t-分布函数
- ◀ seaborn.histplot() 绘制频率/概率直方图



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

7.1 连续均匀分布：离散均匀分布的连续版

概率密度函数

如图 1 所示，连续随机变量 X 在区间 $[a, b]$ 内取得任意一个实数的概率密度函数满足：

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \quad (1)$$

则称 X 区间 $[a, b]$ 上服从**连续均匀分布** (continuous uniform distribution)。这个连续分布常记做 $\text{Uniform}(a, b)$ 或 $U(a, b)$ ，比如 $[0, 1]$ 区间上的均匀分布可以记做 $\text{Uniform}(0, 1)$ 或 $U(0, 1)$ 。

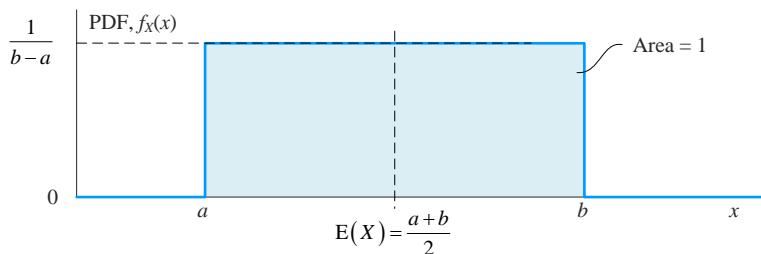


图 1. 随机变量 X 在 $[a, b]$ 上为均匀分布

期望、方差

服从 (1) 连续均匀分布 X 的期望和方差分别为：

$$\mathbb{E}(X) = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)^2}{12} \quad (2)$$

随机数

利用随机数发生器，我们可以获得满足连续均匀分布的随机数。图 2 (a) 所示为满足连续均匀分布随机数的直方图。

图 2 (b) 所示为随机数的**经验累积分布函数** (Empirical Cumulative Distribution Function, ECDF)。不难看出 ECDF 的取值范围为 $[0, 1]$ 。经验分布函数是在所有 n 个样本点上都跳跃 $1/n$ 的阶跃函数。对于某个特定样本，它的 ECDF 为样本中小于或等于该值的样本所占的比例。



我们在本书第 9 章还会提到经验累积分布函数 ECDF。

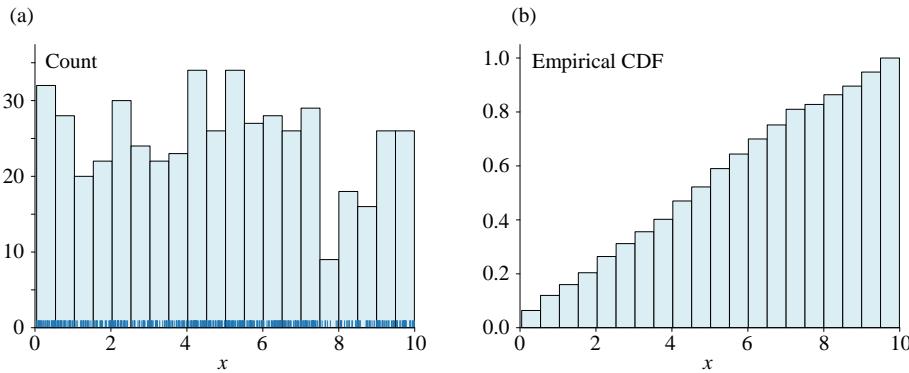


图 2. 满足连续均匀分布的随机数直方图、ECDF



Bk5_Ch07_01.py 代码绘制图 2。

7.2 高斯分布：最重要的概率分布，没有之一

高斯分布 (Gaussian distribution)，也叫**正态分布** (normal distribution)，仿佛是整个纷繁复杂宇宙表象下的终极秩序。实际上，高斯分布是由德国数学家和天文学家**亚伯拉罕·棣莫弗** (Abraham de Moivre) 于 1733 年首先提出。

→ 高斯分布非常重要，“鸢尾花书”中回归分析、主成分分析、高斯朴素贝叶斯、高斯过程、高斯混合模型等等内容都和高斯分布有着密切的联系。本书第 9 ~ 13 章将从不同角度探讨高斯分布。

一元高斯分布

一元高斯分布 (univariate normal distribution) 的概率密度函数为：

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (3)$$

其中， μ 为均值/期望值， σ 为标准差。满足 (3) 的高斯分布常记做 $N(\mu, \sigma^2)$ 。

也就是说，连续随机变量 X 服从 $N(\mu, \sigma^2)$ ，即 $X \sim N(\mu, \sigma^2)$ ，则 X 的期望和方差为：

$$\mathbb{E}(X) = \mu, \quad \text{var}(X) = \sigma^2 \quad (4)$$

图 3 所示为三个不同一元高斯分布 PDF、CDF 图像。可以发现，一元高斯分布 PDF 关于 $x = \mu$ 对称，当 x 远离 μ ，概率密度函数高度迅速下降。

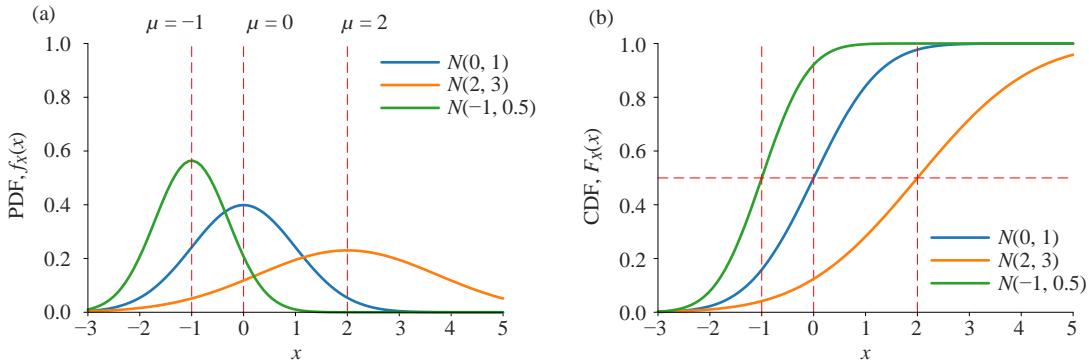


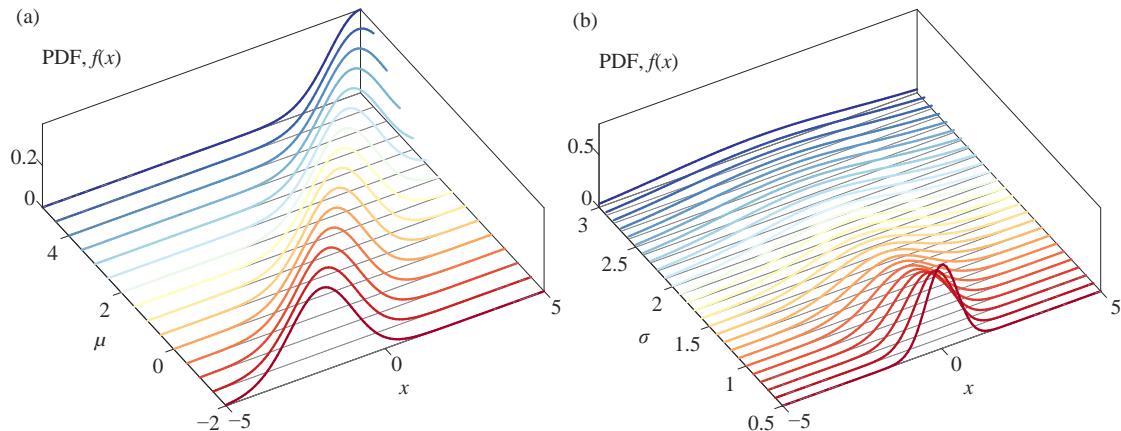
图 3. 三个正态分布 PDF 和 CDF



Bk5_Ch07_02.py 代码绘制图 3。

形状

μ 和 σ 两个参数确定了一元高斯分布 PDF 的位置和形状。如图 4 所示， μ 决定概率密度曲线 $p(x)$ 的位置， σ 影响曲线的胖瘦。特别地，当 $\mu = 0$ ，且 $\sigma = 1$ 时，得到的高斯分布为**标准正态分布** (standard normal distribution)。

图 4. 均值 μ 和标准差 σ 分别对一元正态分布曲线形状影响

二元高斯分布

二元高斯分布 (bivariate Gaussian distribution)，也叫**二元正态分布**，它的概率密度函数解析式如下：

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{1,2}^2}} \times \exp\left\{-\frac{1}{2}\left(\frac{1}{(1-\rho_{1,2}^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{1,2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)\right)\right\} \quad (5)$$

其中， μ_1 和 μ_2 分别为 X_1 和 X_2 的期望值， σ_1 和 σ_2 为 X_1 和 X_2 的标准差， $\rho_{1,2}$ 为两者的线性相关系数。

⚠ 注意，上式中 $\rho_{1,2}$ 取值范围为 $(-1, 1)$ 。

→ 相信大家已经在上式中看到椭圆！这是本书后续重要的线索之一。此外，我们在《数学要素》第 9 章专门介绍过这种椭圆形式。

连续随机变量 (X_1, X_2) 服从上述二元正态分布，记做：

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \underbrace{\begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}}_{\Sigma}\right) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (6)$$

图 5 所示为方差和相关性系数取不同值时，二元正态分布概率密度函数椭圆等高线以及边缘分布形状。注意，图中 $\sigma_{1,1}$ 和 $\sigma_{2,2}$ 代表方差，即标准差的平方。

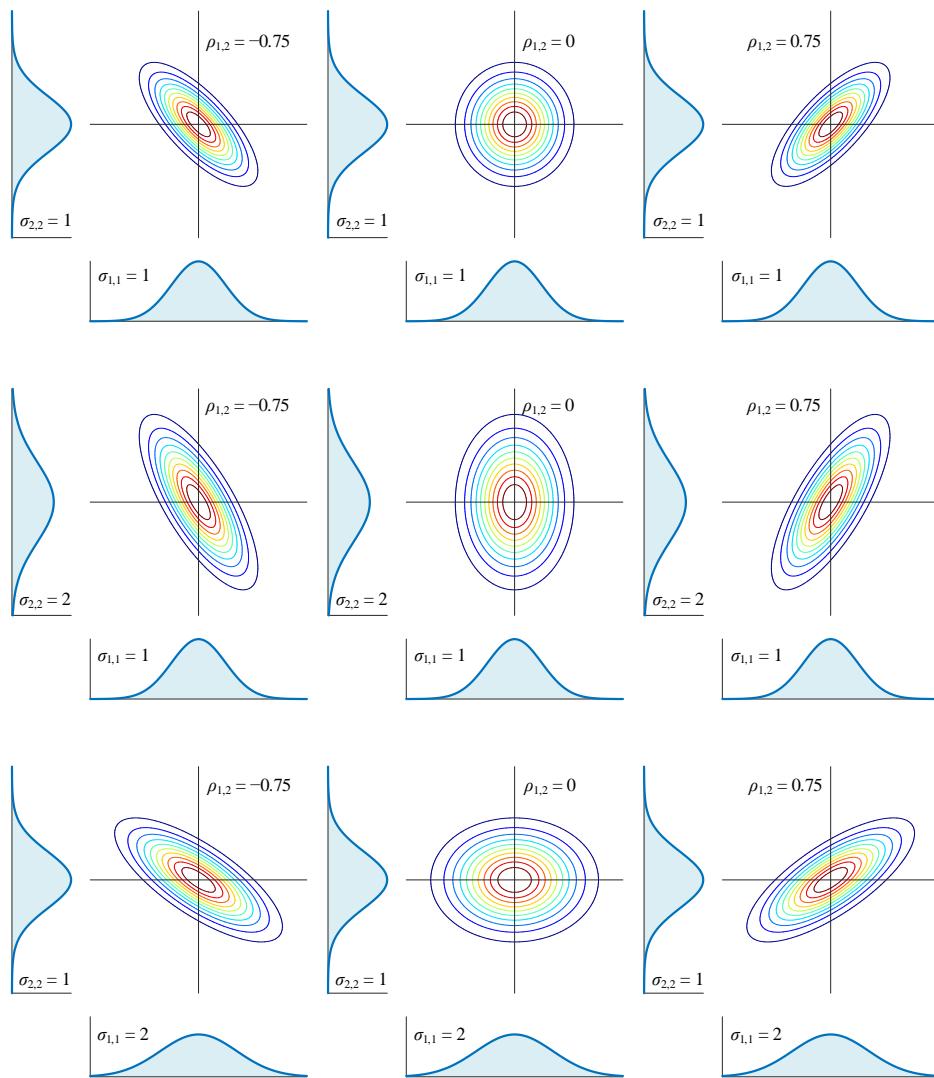


图 5. 方差和相关性系数取不同值时，二元正态分布概率密度函数椭圆等高线形态



本书第 10 章将专门以椭圆为视角讲解二元正态分布。

多元高斯分布

《矩阵力量》第 20 章用如下公式介绍过**多元高斯分布** (multivariate Gaussian distribution)，请大家据此回忆多元高斯分布 PDF 每个不同成分的含义：

$$f_x(x) = \frac{\exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}}$$

(7)

该公式展示了多元高斯分布 PDF 的构建过程：

- $d = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$ | Mahal distance
- $\|z\|$ | z-score
- $z = A^{\frac{-1}{2}} V^T (x - \mu)$ | Translate → rotate → scale
- $\left[A^{\frac{-1}{2}} V^T (x - \mu) \right]^T A^{\frac{-1}{2}} V^T (x - \mu)$ | Eigen decomposition
- $(x - \mu)^T \Sigma^{-1} (x - \mu)$ | Ellipse/ellipsoid
- $f_x(x)$ | Distance → similarity
- \downarrow | Normalization
- \downarrow | Multivariable calculus
- \downarrow | Scaling
- \downarrow | Eigenvalues



本书第 11 章深入讲解多元高斯分布。

拉普拉斯分布：

本节最后简要介绍**拉普拉斯分布** (Laplace distribution)。拉普拉斯分布的概率密度函数为：

$$f_x(x) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right) \quad (8)$$

形式上，拉普拉斯分布和高斯分布很类似，只不过拉普拉斯分布的 PDF 图像在对称轴处存在尖点。很容易发现，参数 μ 决定概率密度分布位置。如图 6 所示，参数 b 决定分布形状。

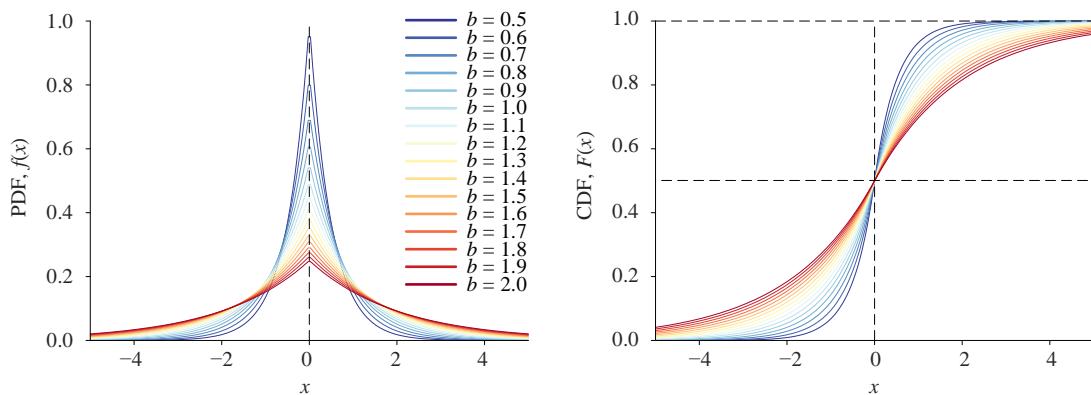


图 6. 拉普拉斯分布的 PDF 和 CDF

如果连续随机变量 X 满足 (8) 拉普拉斯分布, X 期望和方差为:

$$\mathbb{E}(X) = \mu, \quad \text{var}(X) = 2b^2 \quad (9)$$

两个常用的拉普拉斯分布函数为 `scipy.stats.laplace()` 和 `numpy.random.laplace()`。



《数学要素》第 12 章分别讲解过高斯函数和拉普拉斯函数，建议大家回顾。

7.3 逻辑分布：类似高斯分布

一元逻辑分布 (univariate logistic distribution) 的 PDF 为:

$$f_x(x) = \frac{\exp\left(\frac{-(x-\mu)}{s}\right)}{s\left(1+\exp\left(\frac{-(x-\mu)}{s}\right)\right)^2} \quad (10)$$

其中, μ 为位置参数, s 为形状参数。

相比 PDF, 逻辑函数的 CDF 更常用:

$$F_x(x) = \frac{1}{1 + \exp\left(\frac{-(x-\mu)}{s}\right)} \quad (11)$$

图 7 所示为逻辑函数的 PDF 和 CDF 曲线随 b 变化。

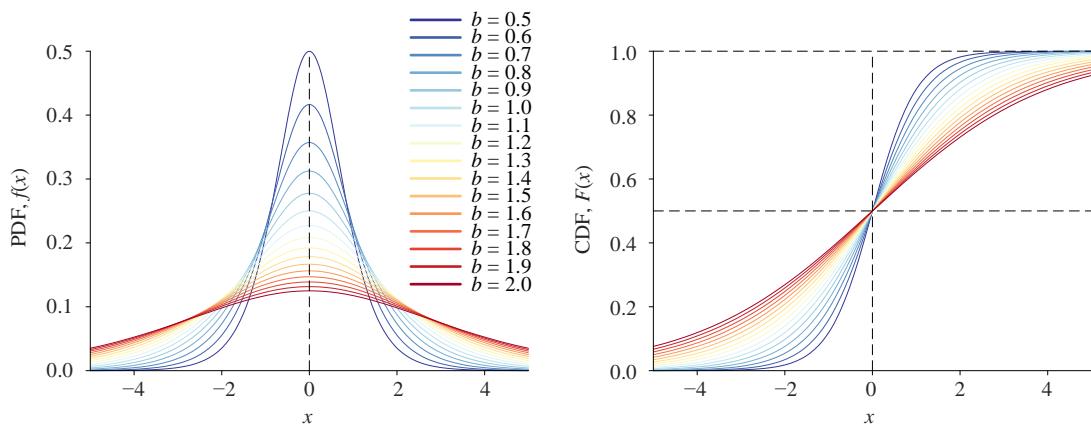


图 7. 逻辑分布 PDF 和 CDF

逻辑分布 vs 高斯分布

大家肯定已经发现，逻辑分布和高斯分布 PDF、CDF 长得很相似。为了比较逻辑函数和高斯函数，我们用标准正态分布 $N(0, 1)$ 的 PDF 和 CDF 图像，而逻辑分布的位置参数 $\mu = 0$ 。特别选取参数 s 使得逻辑分布 PDF 和标准正态分布 PDF 在 $x = 0$ 处高度一致。

如图 8 所示，相比标准正态分布，逻辑分布 PDF 中心部位“稍瘦”，而**厚尾** (fat tail)。厚尾，也叫肥尾，指的是和正态分布相比，尾部分布较厚的分布。下一节介绍的学生 t -分布就是典型的厚尾分布。

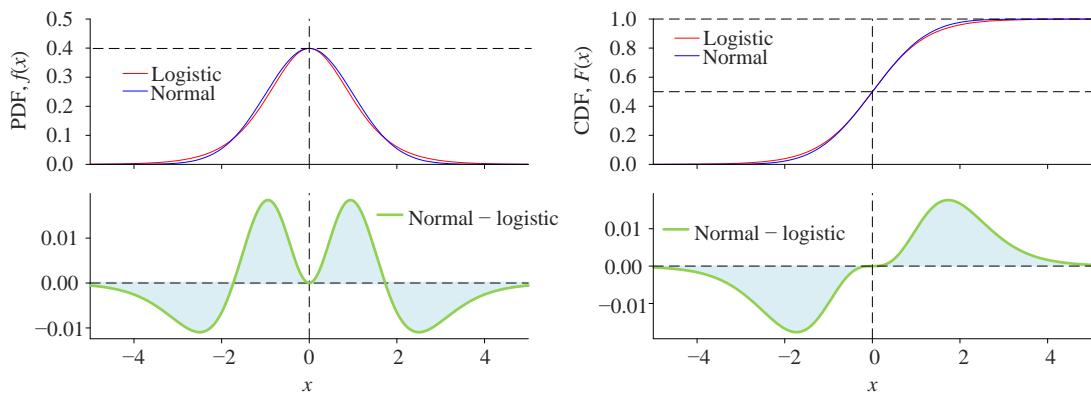
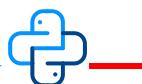


图 8. 比较逻辑函数和高斯函数



Bk5_Ch07_03.py 代码绘制图 7。

7.4 学生 t -分布：厚尾分布

学生 t -分布 (Student's t-distribution) 也称**学生分布**，或 t 分布，是由**戈赛特** (William Sealy Gosset) 于 1908 年提出的，Student 一词源自于他发表论文时用的化名。

学生 t -分布是常用的一类厚尾分布。学生 t -分布多应用于根据小样本数据来估计呈正态分布且方差未知的总体的均值，本书第 17 章将简要介绍相关内容。

一元学生 t -分布的 PDF 为：

$$f_x(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{(\nu+1)}{2}} \quad (12)$$

其中， ν 为**自由度** (number of degrees of freedom 或 df)， $\nu = n - 1$ ， n 为样本数； Γ 是 **Gamma 函数** (Gamma function)。

Gamma 函数

Gamma 函数是从阶乘的概念推广而来的，它将阶乘的概念推广到了实数和复数的范围。

ν 为正整数时，Gamma 方程类似于阶乘表达式，正整数 ν 的 Gamma 函数表达式为：

$$\Gamma(\nu) = (\nu - 1)! \quad (13)$$

ν 取特殊分数，比如 $1/2$ 和 $3/2$ 时， ν 的 Gamma 函数的值：

$$\begin{aligned} \Gamma\left(\frac{1}{2}\right) &= \sqrt{\pi} \\ \Gamma\left(\frac{3}{2}\right) &= \frac{1}{2}\sqrt{\pi} \end{aligned} \quad (14)$$

图 9 所示为 Gamma 函数图像，其中红色 \times 是取正整数时 Gamma 函数的取值。

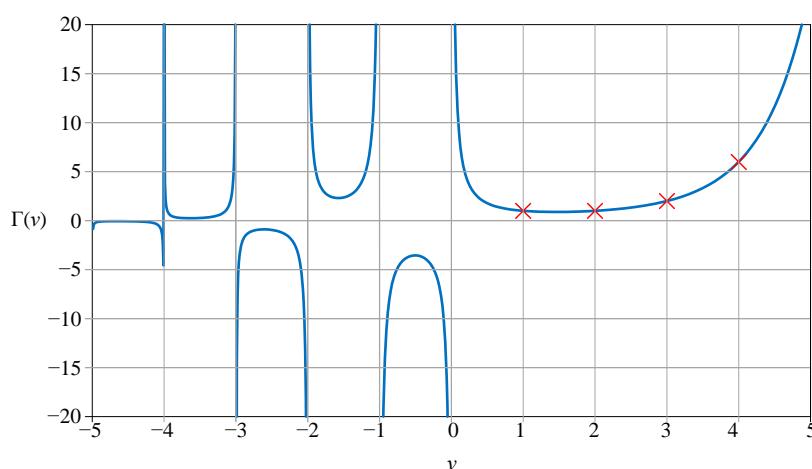


图 9. Gamma 函数图像

一般情况，当 ν 为偶数时，(12) 中系数部分为：

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} = \frac{(\nu-1)(\nu-3)\cdots 5 \cdot 3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots 4 \cdot 2} \quad (15)$$

当 ν 为奇数时：

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi} \cdot \Gamma\left(\frac{\nu}{2}\right)} = \frac{(\nu-1)(\nu-3)\cdots 4 \cdot 2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots 5 \cdot 3} \quad (16)$$

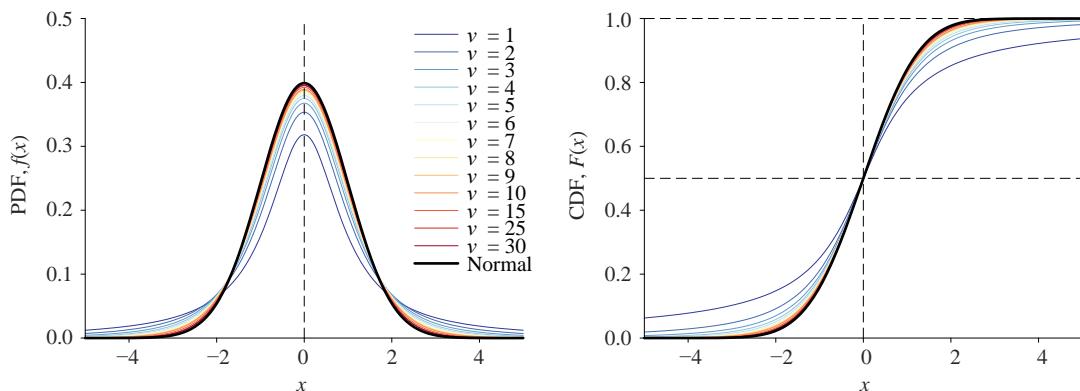
Gamma 函数存在如下递推关系：

$$\Gamma(\nu+1) = \Gamma(\nu) \cdot \nu \quad (17)$$

上式和 ν 取值无关。Gamma 函数在概率分布中具有重要的作用，尤其是在 Gamma 分布、卡方分布、 t 分布、Beta 分布、Dirichlet 分布等定义和性质中都涉及到 Gamma 函数。

自由度

图 10 所示为 ν 从 1 变化到 30 时，学生 t -分布 PDF 和 CDF 图像。图 10 中黑色的曲线对应正态分布。当自由度 ν 不断提高时，厚尾现象逐渐消失，学生 t -分布逐渐接近标准正态分布（黑色）。很明显，学生 t -分布的偏度为 0。

图 10. 学生 t -分布 PDF 和 CDF 随自由度变化

Bk5_Ch07_04.py 代码绘制图 10。

多元学生 t -分布

类似(7)给出的多元高斯分布，多元学生 t -分布的概率密度函数为：

$$f_x(\mathbf{x}) = \frac{\Gamma[(\nu+D)/2]}{\Gamma(\nu/2)\nu^{D/2}\pi^{D/2}|\boldsymbol{\Sigma}_t|^{1/2}} \left[1 + \underbrace{\frac{1}{\nu} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}_t^{-1} (\mathbf{x} - \boldsymbol{\mu})}_{\text{Ellipse}} \right]^{-(\nu+D)/2} \quad (18)$$

其中， ν 为自由度， D 为维数。相信大家在上式中也看到了椭圆。

上式中 $\boldsymbol{\Sigma}_t$ 和多元高斯分布的协方差矩阵关系为：

$$\boldsymbol{\Sigma}_t = \frac{\nu}{\nu-2} \boldsymbol{\Sigma} \quad (19)$$

7.5 对数正态分布：源自正态分布

定义

如果随机变量 X 的对数 $\ln X$ 服从正态分布，则 X 服从**对数正态分布** (logarithmic normal distribution)。

对于 $x > 0$ ，对数正态分布的 PDF 为：

$$f_x(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \quad (20)$$

其中， μ 是 X 对数的平均值， σ 是 X 对数的标准差。

如果 X 满足(20)的对数正态分布，则 X 期望和方差为：

$$E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right), \quad \text{var}(X) = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2) \quad (21)$$

图像

图 11 给出对数正态分布的图像。对数正态分布的最大特点是右偏，即正偏。对于右偏的对数正态分布，其平均值大于其众数。



大家将会在《数据有道》一册看到对数正态分布的应用。

⚠ 再次强调，对数正态分布的随机变量取值只能为正值。

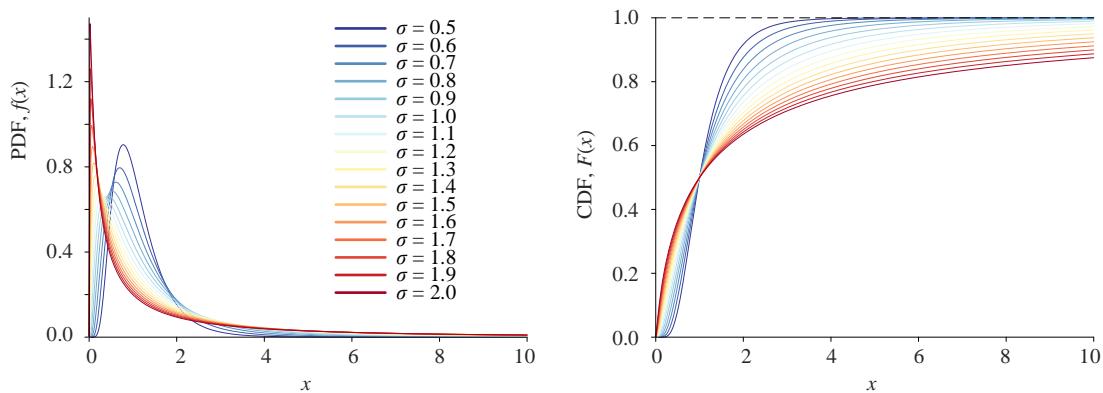


图 11. 对数正态分布的 PDF 和 CDF

图 12 对比正态分布和对数正态分布。

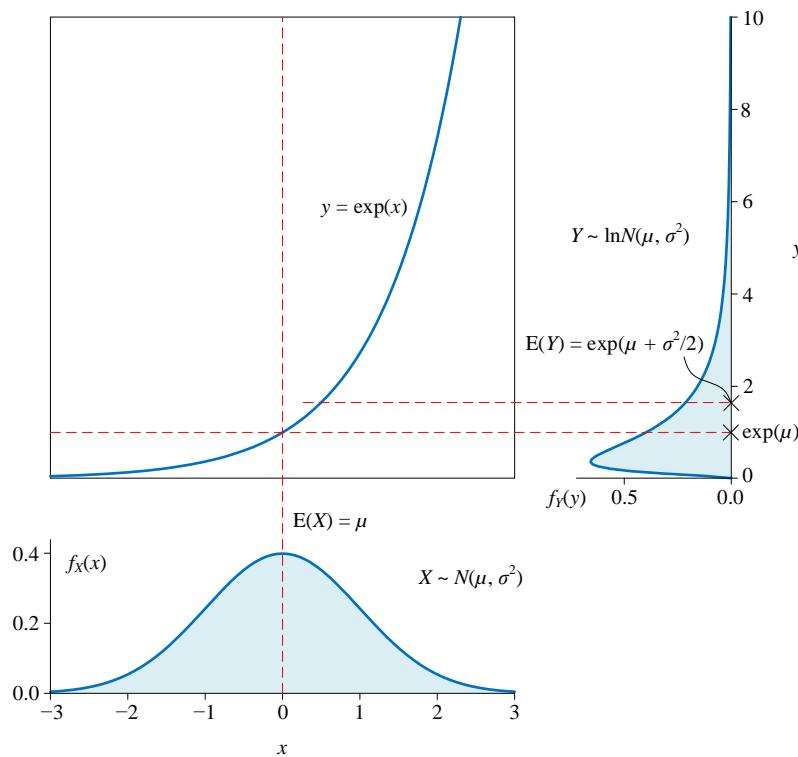
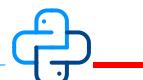


图 12. 比较正态分布和对数正态分布



Bk5_Ch07_05.py 代码绘制图 11。Bk5_Ch07_06.py 代码绘制图 12

7.6 指数分布：泊松分布的连续随机变量版

定义

指数分布 (exponential distribution) 和本章第 5 章介绍的泊松分布息息相关。

与泊松分布相比，指数分布重要特点是随机变量连续。而泊松分布是针对随机事件发生次数定义的，发生次数是离散的。

指数分布的概率密度函数为：

$$f_x(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (22)$$

指数分布的期望和方差分别为：

$$\mathbb{E}(X) = \frac{1}{\lambda}, \quad \text{var}(X) = \frac{1}{\lambda^2} \quad (23)$$

图像

图 13 所示为 λ 取不同值时，指数分布 PDF 和 CDF 图像。

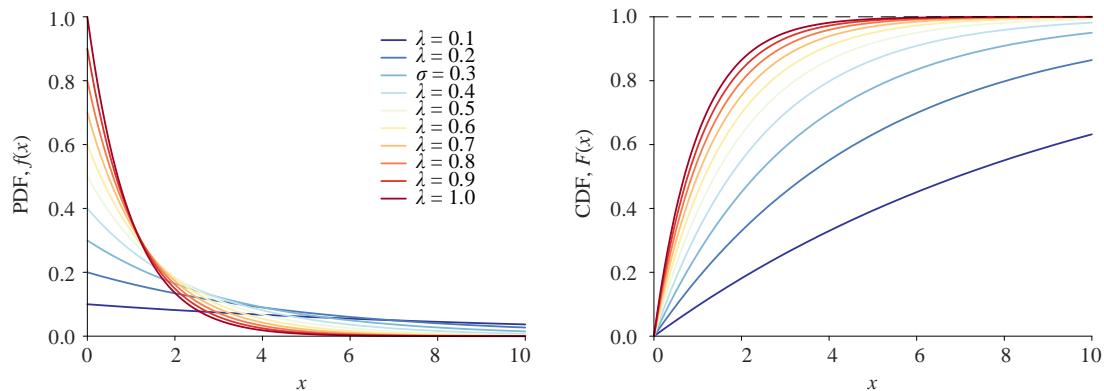


图 13. λ 取不同值时，指数分布 PDF 和 CDF 图像



Bk5_Ch07_07.py 代码绘制图 13。

7.7

卡方分布：若干 IID 标准正态分布平方和

定义

卡方分布 (chi-square distribution 或 χ^2 -distribution) 先是德国统计学家**赫尔默特** (Friedrich Robert Helmert) 在 1875 年提出。

若 n 个相互独立的随机变量 Z_1, Z_2, \dots, Z_k 均服从标准正态分布，即：

$$Z_i \sim N(0,1), \quad \forall i = 1, \dots, k \quad (24)$$

这 n 个随机变量的平方和构成一个新的随机变量 X ， X 服从自由度为 k 的卡方分布：

$$X = \sum_{i=1}^k Z_i^2 \sim \chi_k^2 \quad (25)$$

其中， k 称为自由度。自由度为 k 的卡方分布一般标记为 χ_k^2 。

如果随机变量 X 满足 (25) 的卡方分布， X 的期望值和方差为：

$$E(X) = k, \quad \text{var}(X) = 2k \quad (26)$$

图像

如图 14 所示，卡方分布的值均为正值，且呈现右偏态，随着自由度 n 的增大，卡方分布趋近于正态分布。当自由度大于 30 时，已经非常类似于正态分布。

不知道大家看到 (25)，是否想到马氏距离的平方？



我们将在本书第 23 章讲解马氏距离时用到卡方分布。

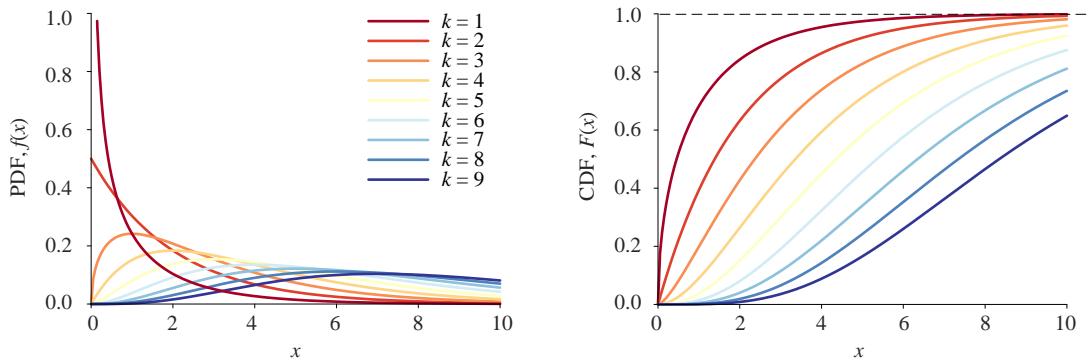
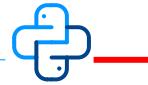


图 14. 卡方分布 PDF 和 CDF



Bk5_Ch07_08.py 代码绘制图 14。

7.8 F-分布：和两个服从卡方分布的独立随机变量有关

定义

F -分布是两个服从卡方分布的独立随机变量各除以其自由度后的比值的抽样分布。

如果随机变量 X 满足参数为 d_1 和 d_2 的 F -分布，记做 $X \sim F(d_1, d_2)$ 。随机变量 X 为：

$$X = \frac{S_1/d_1}{S_2/d_2} \quad (27)$$

其中，随机变量 S_1 和 S_2 分别服从自由度为 d_1 、 d_2 的卡方分布。

如果 $X \sim F(d_1, d_2)$ ， X 的 PDF 为：

$$f_x(x; d_1, d_2) = \frac{1}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1-1}{2}} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{(d_1+d_2)}{2}} \quad (28)$$

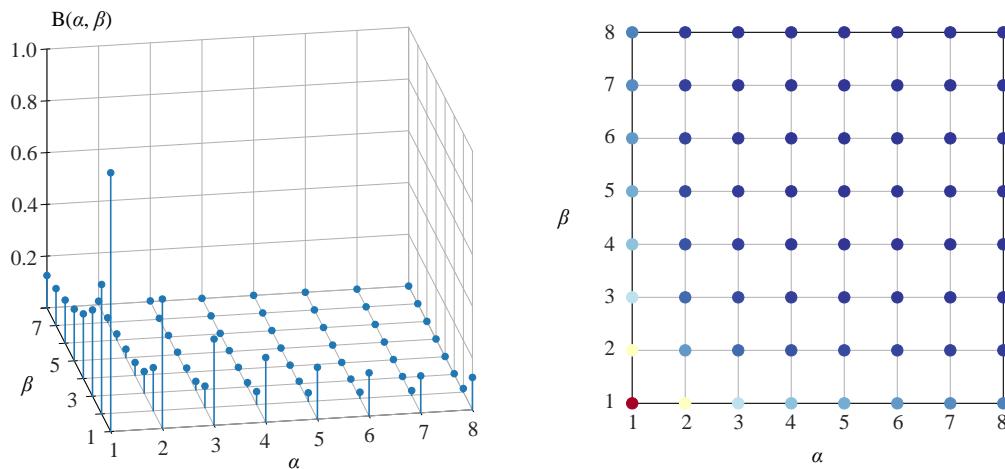
其中， $B()$ 叫做 Beta 函数。Beta 函数和 Gamma 函数的关系为：

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha) \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (29)$$

请大家特别注意上式中的积分式，我们将在本书第 21 章讲解贝叶斯推断时用到这个积分式。

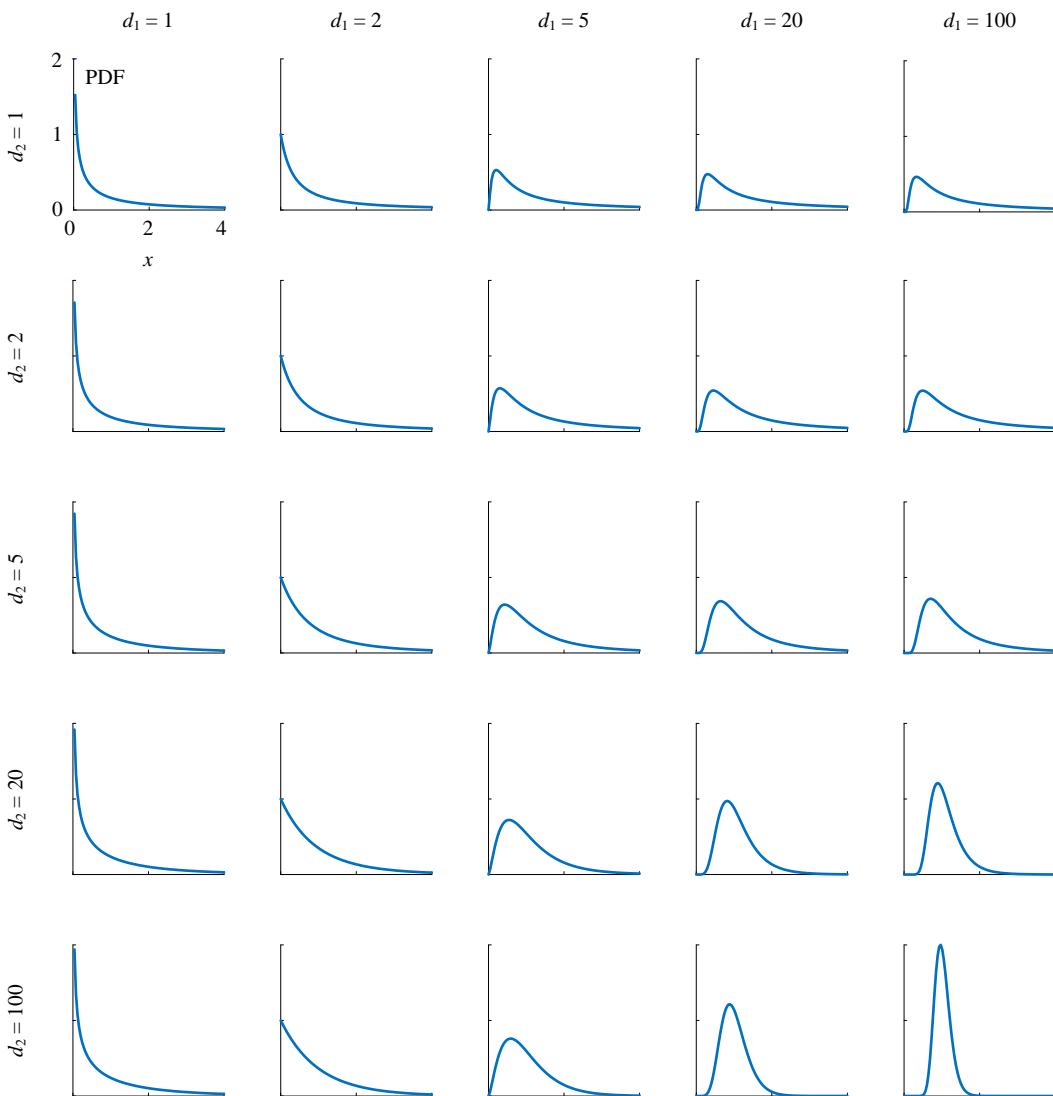
图像

图 15 所示为 $B(\alpha, \beta)$ 函数取值随 α 和 β 变化的火柴梗图、三维散点图。下一节的 Beta 分布中也会用到 $B(\alpha, \beta)$ 函数。

图 15. $B(\alpha, \beta)$ 函数取值火柴梗图、三维散点图

如图 16 所示， F -分布是一种非对称分布，且 d_1 、 d_2 的位置不可随意互换。

在“鸢尾花书”中， F -分布将用在《数据有道》中的**方差分析** (analysis of variance, ANOVA) 和线性回归显著性检验。

图 16. F 分布 PDF 形状随 d_1 和 d_2 变化

Bk5_Ch07_09.py 代码绘制图 16。

7.9 Beta 分布：概率的概率

贝叶斯推断 (Bayesian inference) 是数据科学和机器学习重要的数学工具，而 Beta 分布在贝叶斯推断中扮演重要角色。

定义

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

Beta 分布定义在 $(0, 1)$ 或 $[0, 1]$ 区间的连续概率分布，它有两个参数 α, β 。Beta(α, β) 分布的概率密度函数为：

$$f_x(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (30)$$

其中， $x^{\alpha-1} (1-x)^{\beta-1}$ 决定 PDF 曲线的形状。

大家可能已经注意到，这个 PDF 概率密度曲线有两个区间，原因是 α, β 当取不同值时 x 的取值范围不同。举个例子，当 α, β 均为 0.1 时，Beta 分布的定义域为 $(0, 1)$ 。

相信大家已经在上述解析式中看到了 $B(\alpha, \beta)$ 函数。利用 $B(\alpha, \beta)$ ，(30) 可以写成：

$$f_x(x; \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)} \quad (31)$$

而 $B(\alpha, \beta)$ 是让 $x^{\alpha-1} (1-x)^{\beta-1}$ 成为概率密度函数的归一化因子。白话说， $B(\alpha, \beta)$ 让 PDF 曲线和横轴围成的图形面积为 1。

如果 α, β 都是大于 1 的正整数， $B(\alpha, \beta)$ 可以展开写成：

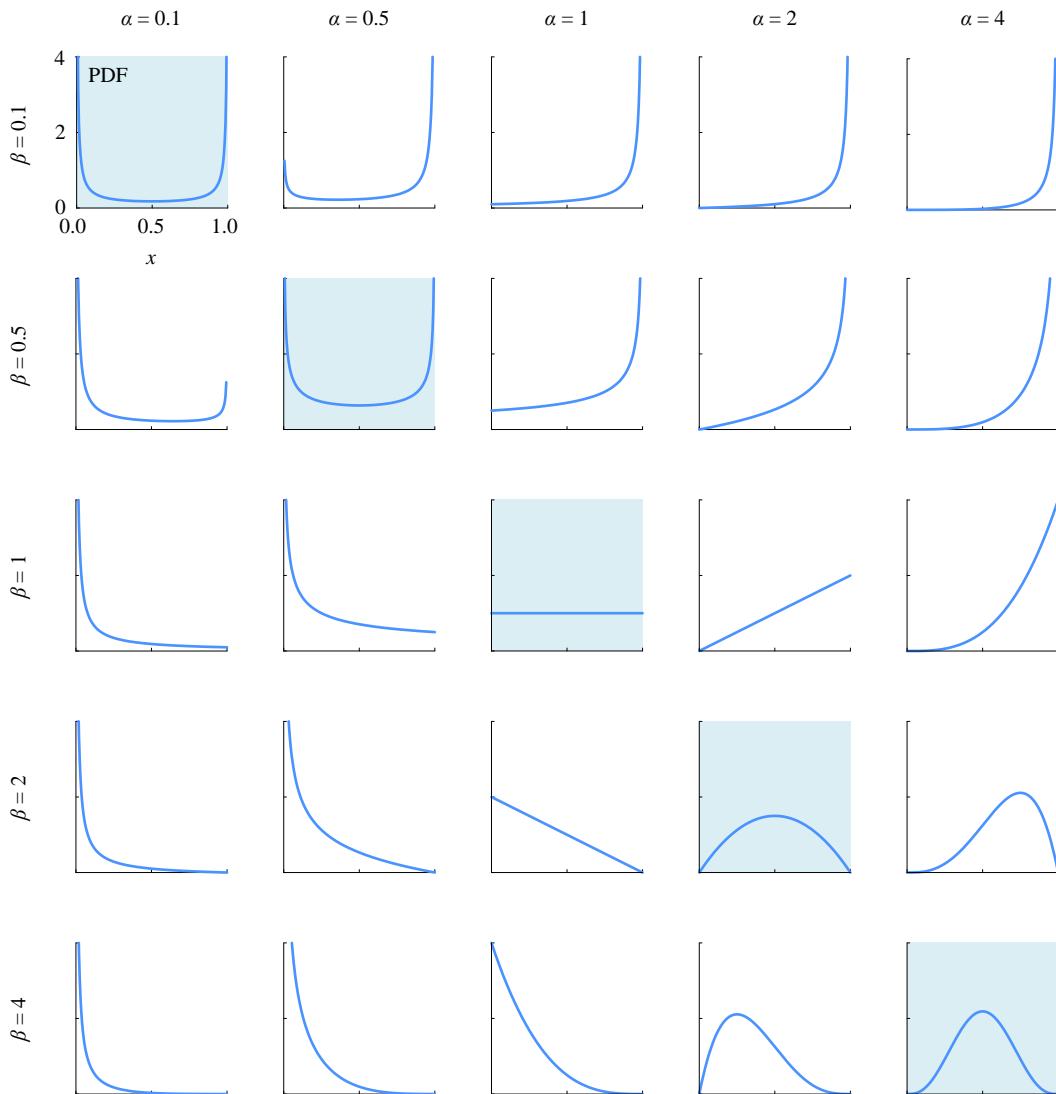
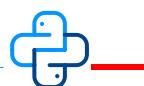
$$B(\alpha, \beta) = \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} \quad (32)$$

图像

图 17 所示为参数 α, β 取不同值时 Beta 分布 PDF 图像。

容易发现 Beta(α, β) 分布实际上代表了一系列分布。举个例子，连续均匀分布 $U(0, 1)$ 便是 Beta($1, 1$)。

请大家特别注意图 17 对角线上的图像，即 $\alpha = \beta$ ，这些 PDF 图像对称，对应的分布相当于 Beta(α, α)。本书第 21 章将用到 Beta(α, α) 这个分布。

图 17. 参数 α 、 β 取不同值时 $\text{Beta}(\alpha, \beta)$ 分布 PDF 图像

Bk5_Ch07_10.py 绘制图 17。代码还绘制 Beta(α, β) 分布的 CDF 图像。



在 Bk5_Ch07_10.py 基础上，我们用 Streamlit 制作了一个应用，大家可以改变 $\text{Beta}(\alpha, \beta)$ 两个参数值，观察 PDF 曲线变化。请大家参考 Streamlit_Bk5_Ch07_10.py。此外，请大家选取前文某个概率分布，做一个类似的 App。

众数 vs 期望

如果 X 服从 $\text{Beta}(\alpha, \beta)$ 分布， X 的期望为：

$$\mathbb{E}(X) = \frac{\alpha}{\alpha + \beta} \quad (33)$$

我们常常用到的是 $\text{Beta}(\alpha, \beta)$ 分布的众数：

$$\frac{\alpha - 1}{\alpha + \beta - 2}, \quad \alpha, \beta > 1 \quad (34)$$

众数是概率密度函数曲线最大值所在位置。这一点在本书后文的贝叶斯推断格外重要，请大家注意。

推导期望

推导 $\text{Beta}(\alpha, \beta)$ 的期望其实很容易，我们甚至不需要积分。

连续随机变量 X 的期望为：

$$\mathbb{E}(X) = \int_x x \cdot f_X(x) dx \quad (35)$$

将 $\text{Beta}(\alpha, \beta)$ 的概率密度函数代入上式，得到：

$$\begin{aligned} \mathbb{E}(X) &= \int_x x \cdot \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} dx \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_x x^\alpha (1-x)^{\beta-1} dx}_{\text{Beta}(\alpha+1, \beta)} \end{aligned} \quad (36)$$

容易看出来，上式中积分部分可以整理成为 $\text{Beta}(\alpha + 1, \beta)$ 分布的 PDF 解析式。缺的就是归一化系数。

补充这个归一化系数，上式可以写成：

$$\begin{aligned} \mathbb{E}(X) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \underbrace{\int_x \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} x^\alpha (1-x)^{\beta-1} dx}_{=1} \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \end{aligned} \quad (37)$$

根据 Gamma 函数的递推关系 $\Gamma(\nu+1) = \Gamma(\nu) \cdot \nu$ ，上式进一步整理为：

$$\begin{aligned} \mathbb{E}(X) &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha) \cdot \alpha \cdot \Gamma(\beta)}{\Gamma(\alpha + \beta) \cdot (\alpha + \beta)} \\ &= \frac{\alpha}{\alpha + \beta} \end{aligned} \quad (38)$$

方差、标准差

Beta(α, β) 的方差为：

$$\text{var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \quad (39)$$

Beta(α, β) 的标准差为方差的平方根：

$$\text{std}(X) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \quad (40)$$

为了方便和下文的 Dirichlet 分布对照，令

$$\alpha_0 = \alpha + \beta \quad (41)$$

Beta(α, β) 的可以进一步写成：

$$\begin{aligned} \text{var}(X) &= \frac{\alpha(\alpha_0 - \alpha)}{\alpha_0^2(\alpha_0 + 1)} \\ &= \frac{\alpha \left(1 - \frac{\alpha}{\alpha_0}\right)}{\alpha_0 + 1} \end{aligned} \quad (42)$$

7.10 Dirichlet 分布：多元 Beta 分布

Dirichlet 分布也叫狄利克雷分布，它本质上是**多元 Beta 分布** (multivariate Beta distribution)。Dirichlet 分布常作为贝叶斯统计的先验概率。

Dirichlet 分布概率密度函数为：

$$f_{x_1, \dots, x_K}(x_1, \dots, x_K; \alpha_1, \dots, \alpha_K) = \frac{1}{B(\alpha_1, \dots, \alpha_K)} \prod_{i=1}^K x_i^{\alpha_i - 1}, \quad \sum_{i=1}^K x_i = 1 \quad (43)$$

注意， x_i ($i = 1, 2, \dots, K$) 的取值范围为 $[0, 1]$ ，而且它们的和为 1。这个分布常记做 $\text{Dir}(\boldsymbol{\alpha})$ 或 $\text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_K)$ 。本书后文在贝叶斯推断中，会用 θ 代替 x 。

K 元 $B()$ 函数的定义为：

$$B(\alpha_1, \dots, \alpha_K) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i\right)} \quad (44)$$

举个例子

当 $K = 3$ 时， x_1 、 x_2 、 x_3 满足：

$$x_1 + x_2 + x_3 = 1 \quad (45)$$

并且， x_1 、 x_2 、 x_3 都在区间 $[0, 1]$ 内。显然， x_1 、 x_2 、 x_3 在一个平面上。

白话说， $x_1 + x_2 + x_3 = 1$ 好比三维空间撑起的一张“画布”，概率密度等高线则必须画在这张画布上。

本节后文将采用五种可视化方案展示 Dirichlet 分布概率密度函数。如图 18 所示，这五种可视化方案主要分成两大类。由于 (45) 等式关系，给定 x_1 、 x_2 ，则 x_3 确定。因此，我们可以用图 18 (a) 的 x_1x_2 平面展示 Dirichlet 分布 PDF 图像。

此外，我们还可以用图 18 (b) 所示的可视化方案。这实际上是**重心坐标系** (barycentric coordinate system)。



“鸢尾花书”《可视之美》专门讲解过重心坐标系，请大家参考。

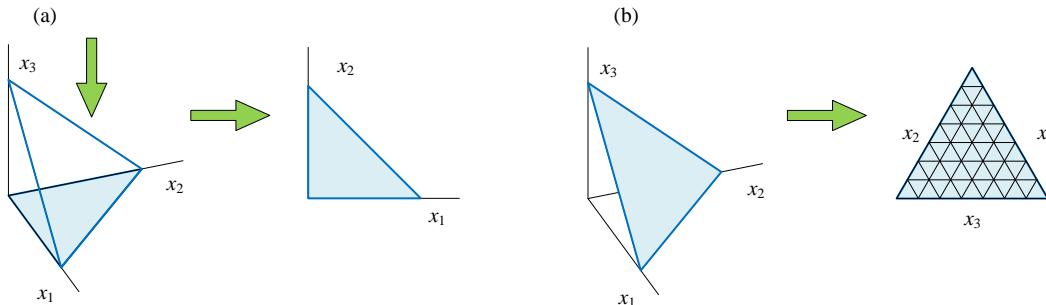


图 18. 可视化方案原理

Dirichlet 分布非常重要，因此我们下文用图 19~图 23 五种可视化方案展示 Dirichlet 分布的分布特征。

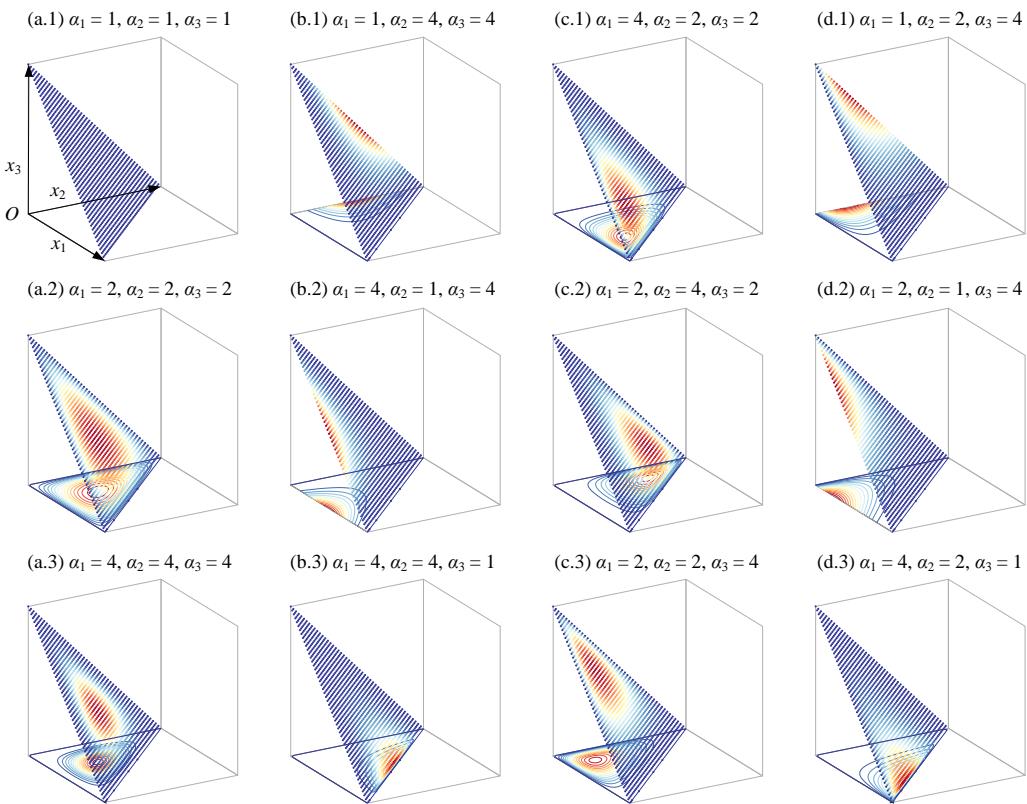
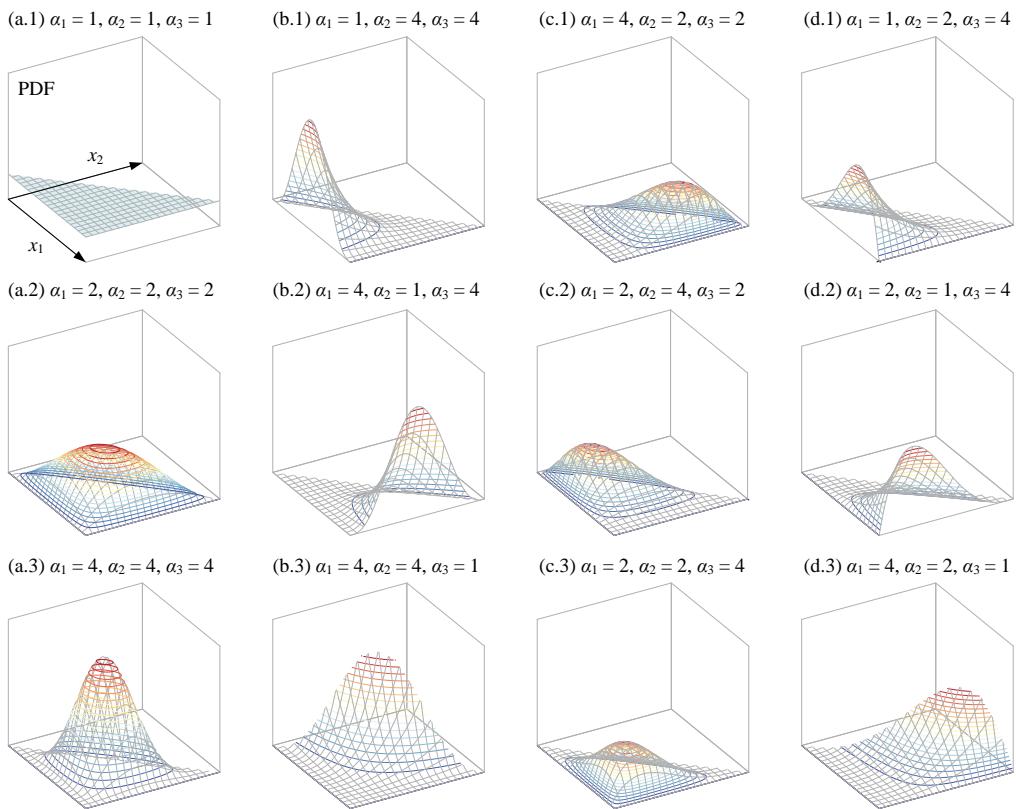


图 19. 用涂色三维散点可视化 Dirichlet 分布图像

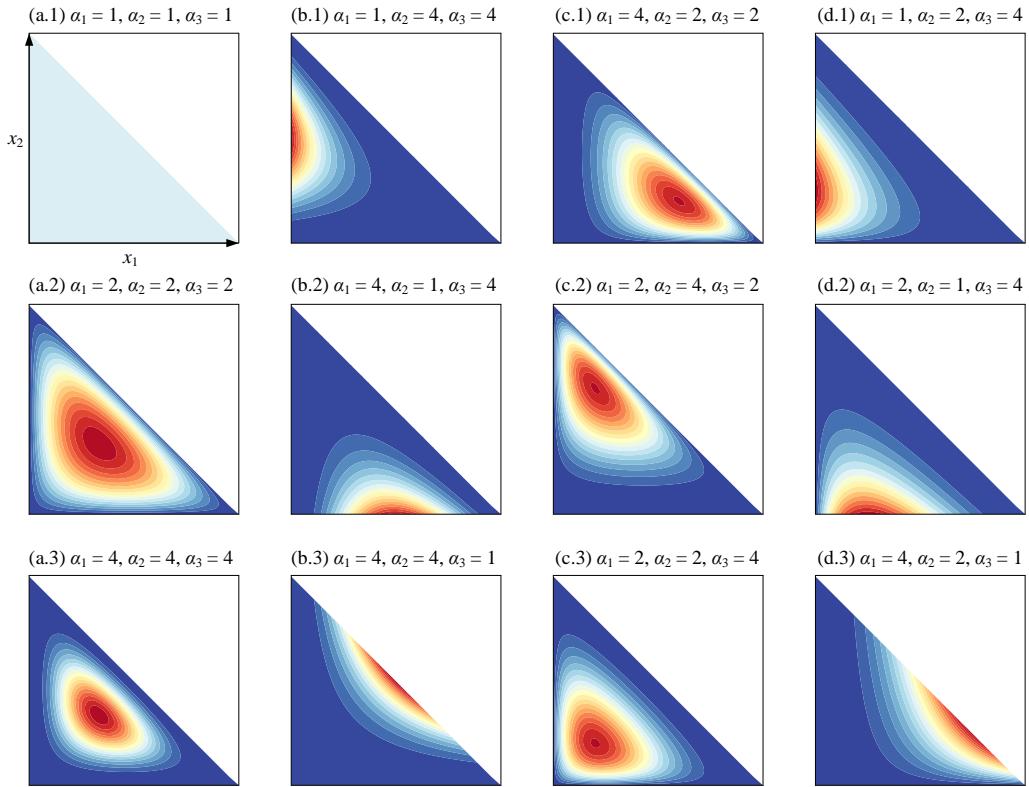
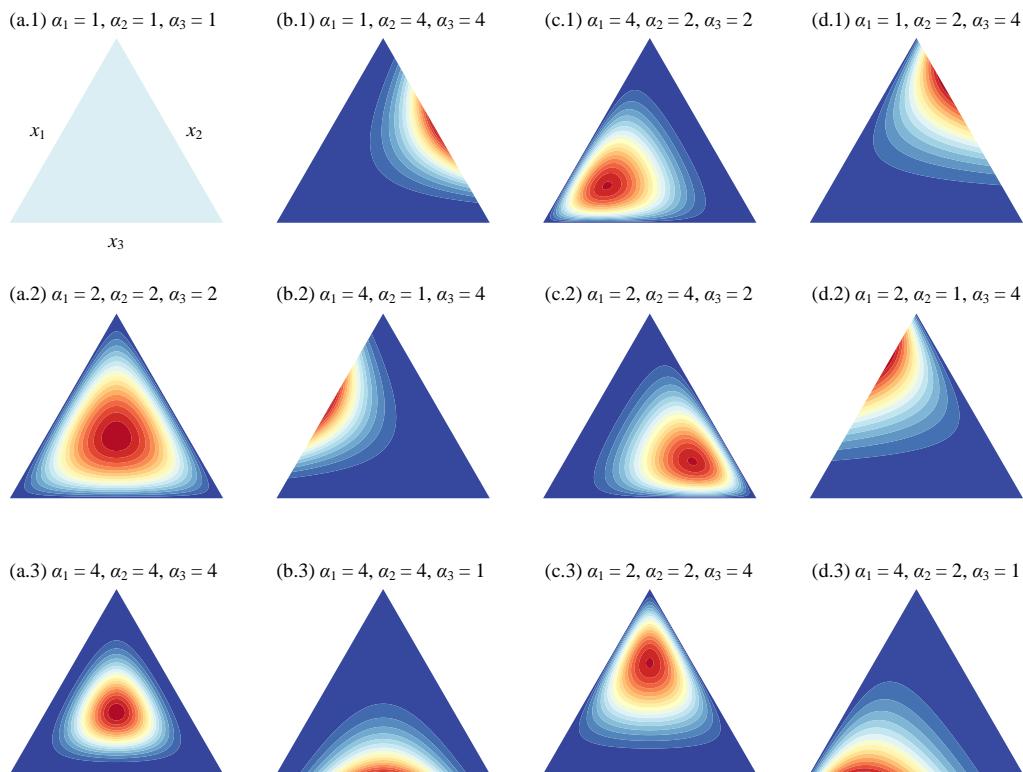


本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 20. 基于 x_1x_2 平面的 Dirichlet 分布 PDF 三维等高线， z 轴为 PDF 取值图 21. x_1x_2 平面上的 Dirichlet 分布 PDF 等高线

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 22. 重心坐标系中的 Dirichlet 分布 PDF 等高线

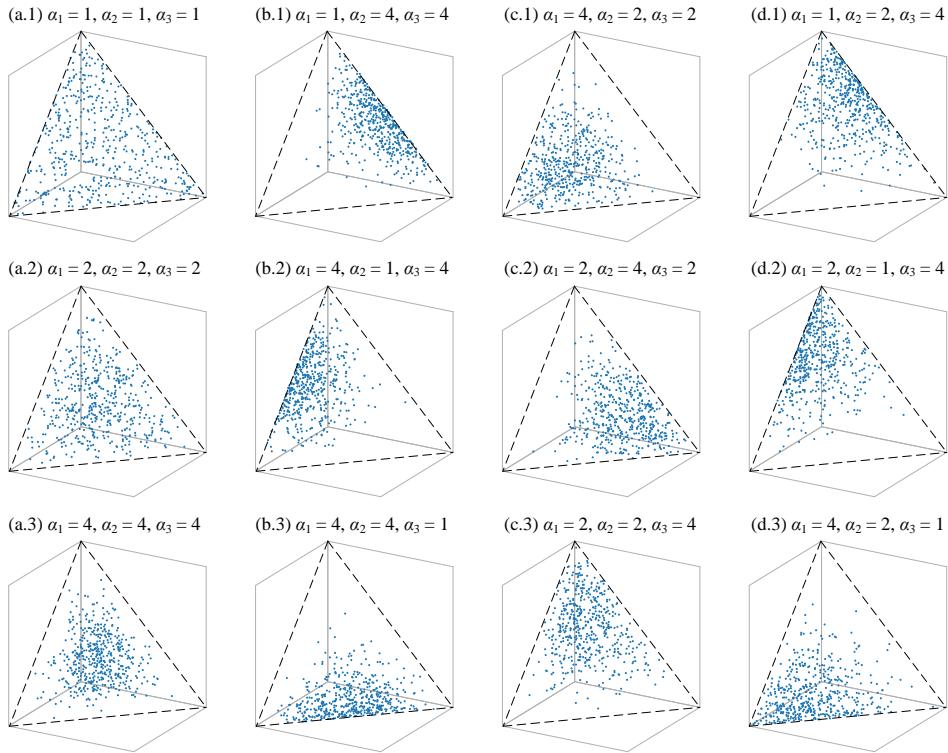


图 23. 满足 Dirichlet 分布的随机数

边缘分布

Dirichlet 分布的边缘分布服从 Beta 分布：

$$X_i \sim \text{Beta}(\alpha_i, \alpha_0 - \alpha_i) \quad (46)$$

其中：

$$\alpha_0 = \sum_{i=1}^K \alpha_i \quad (47)$$

以图 19 中 (d) 组为例，三个 Dirichlet 分布的边缘分布 PDF 如图 24 所示。

X_i 的期望为：

$$E(X_i) = \frac{\alpha_i}{\sum_{k=1}^K \alpha_k} = \frac{\alpha_i}{\alpha_0} \quad (48)$$

X_i 的众数为：

$$\frac{\alpha_i - 1}{\sum_{k=1}^K \alpha_k - K} = \frac{\alpha_i - 1}{\alpha_0 - K}, \quad \alpha_i > 1 \quad (49)$$

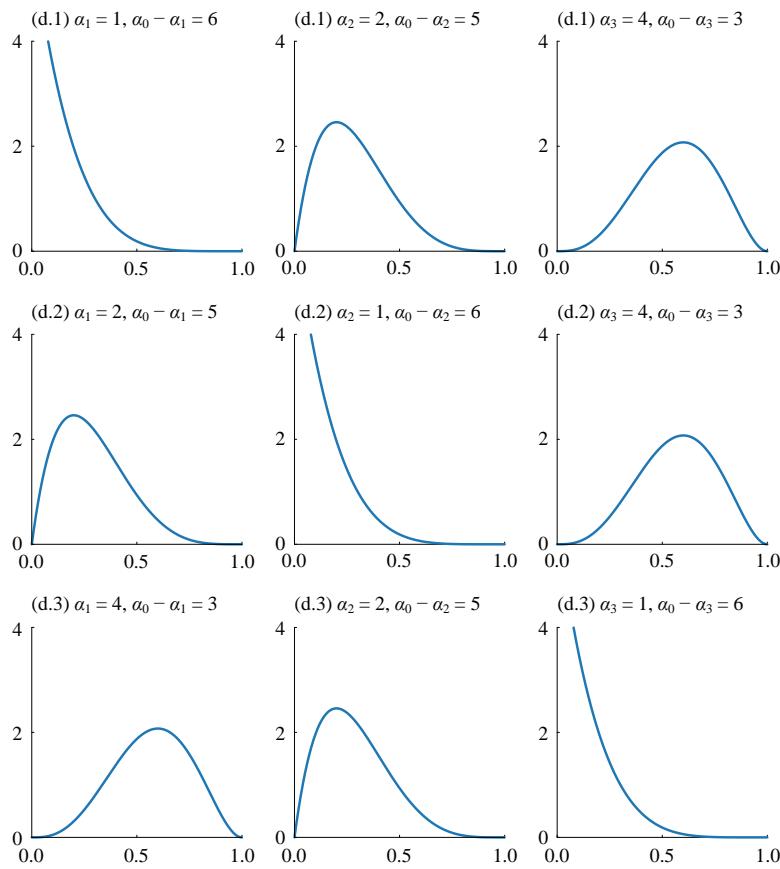
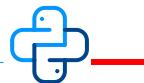


图 24. 三个 Dirichlet 分布的边缘分布



Bk5_Ch07_11.py 绘制图 19、图 20、图 21、图 23、图 24。



在 Bk5_Ch07_11.py 基础上，我们用 Streamlit 制作了一个应用，大家可以改变 $\text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ 三个参数值，观察 PDF 曲面变化。请大家参考 Streamlit_Bk5_Ch07_11.py。



《统计至简》一册整体来看，高斯分布更为重要，但是它不是本章的重点。这一章最重要的分布有两个——Beta 分布、Dirichlet 分布。它俩分别对应本书第 5 章的二项分布、多项分布。这四个分布在本书后续贝叶斯推断中将扮演重要角色。

8

Conditional Expectation and Variance

条件概率

离散、连续随机变量的条件期望、条件方差



每一种科学，只要达到一定程度的成熟，就会自动成为数学的一部分。

Every kind of science, if it has only reached a certain degree of maturity, automatically becomes a part of mathematics.

—— 大卫·希尔伯特 (David Hilbert) | 德国数学家 | 1862 ~ 1943



- ◀ `matplotlib.pyplot.errorbar()` 绘制误差棒
- ◀ `matplotlib.pyplot.stem()` 绘制火柴梗图
- ◀ `numpy.mean()` 计算均值
- ◀ `numpy.sqrt()` 计算平方根
- ◀ `numpy.std()` 计算标准差，默认分母为 n，不是 n - 1
- ◀ `numpy.var()` 计算方差，默认分母为 n，不是 n - 1
- ◀ `seaborn.heatmap()` 绘制热图

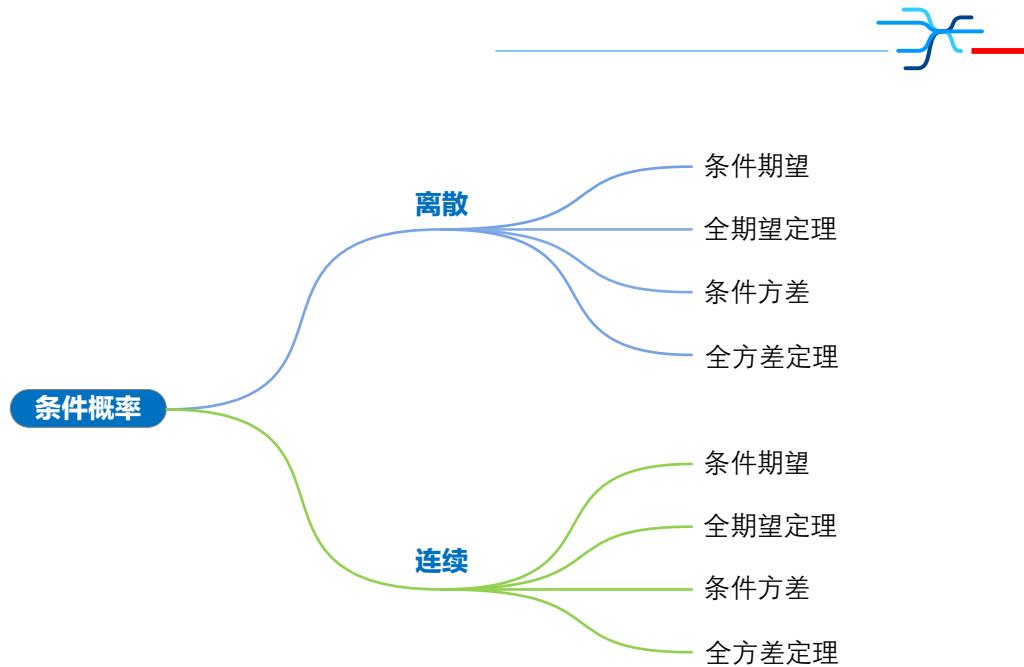
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



8.1 离散随机变量：条件期望

条件期望 (conditional expectation 或 conditional expected value), 或条件均值 (conditional mean), 是一个随机变量的相对于一个条件概率分布的期望。换句话说, 这是给定的一个或多个其他随机变量值的条件下, 某个特定随机变量的期望。

类似地, 条件方差 (conditional variance) 与一般方差的定义几乎一致。计算条件方差时, 只不过将期望换成了条件期望, 并将概率换成了条件概率而已。

条件期望和条件方差这两个概念在数据科学、机器学习算法中格外重要, 本章分别讲解离散随机变量和随机变量的条件期望和条件方差。



本书第 12 章则专门介绍高斯条件概率。

大家应该已经看到, 本章期望、方差交替出现, 为了帮助大家阅读, 我们用给期望、方差涂了不同颜色。

什么是条件期望?

条件期望其实很好理解。比如, 一个笼子里 10 只动物, 其中 6 只鸡 (60%)、4 只兔 (40%)。如图 1 所示, 分别只考虑鸡, 只考虑兔, 这就是“条件”。

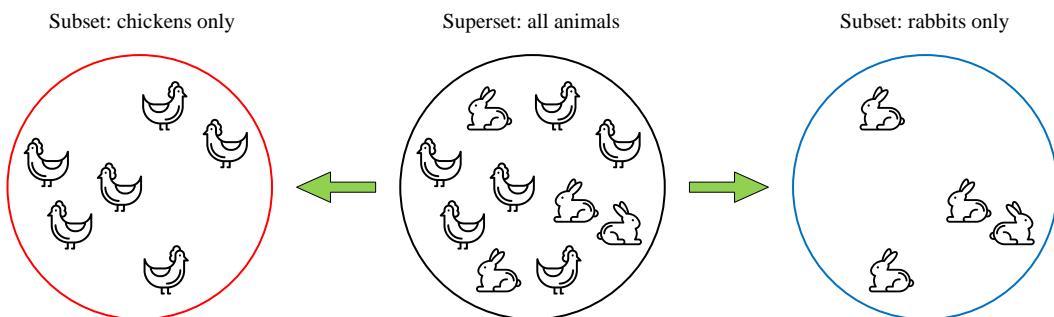


图 1. 解释条件

如图 2 所示, 鸡的平均体重为 2 公斤, 这个数值就是条件期望。再举个例子, 兔子的平均体重为 4 公斤, 这也是条件期望。

本书后续会用鸢尾花数据做例子给大家继续讲解条件期望。

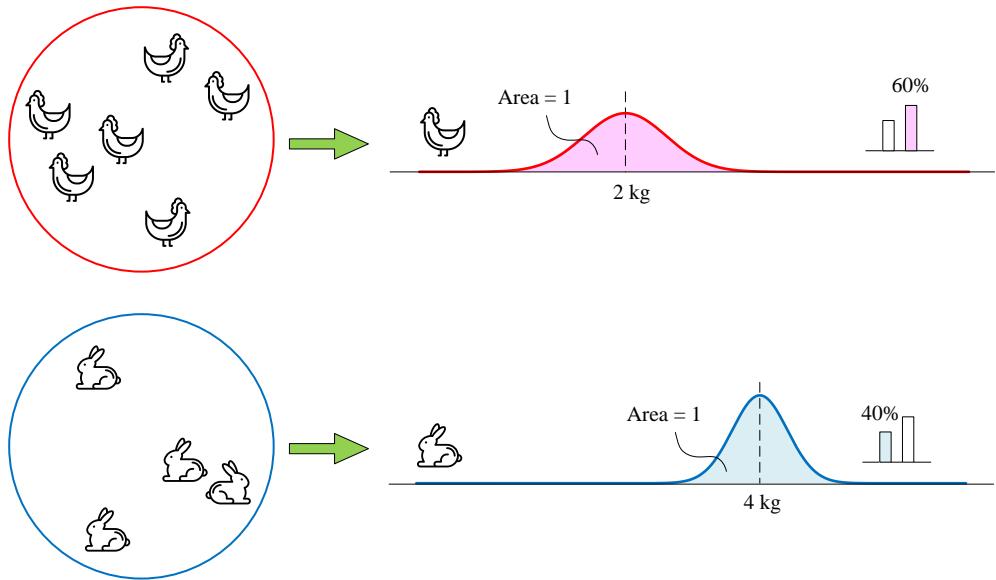


图 2. 解释条件期望

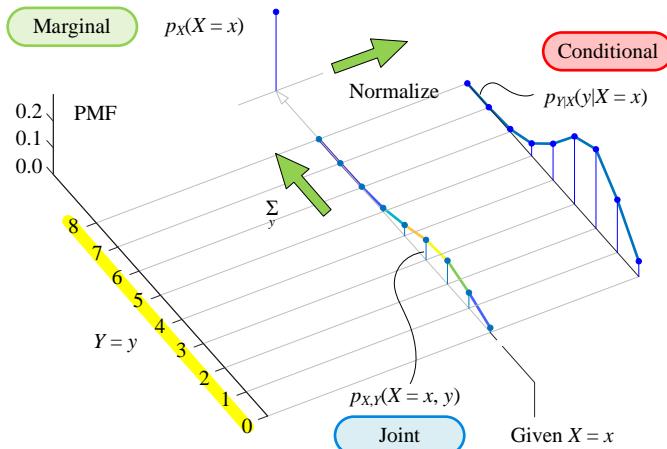
条件期望 $E(Y|X = x)$

如果 X 和 Y 均为离散随机变量，给定 $X = x$ 条件下， Y 的条件期望 $E(Y|X = x)$ (conditional mean of Y given $X = x$) 定义为：

$$\begin{aligned} E(Y \mid X = x) &= \sum_y \underbrace{y \cdot p_{Y|X}(y|x)}_{\text{Conditional}} \\ &= \sum_y y \cdot \underbrace{\frac{p_{X,Y}(x,y)}{p_X(x)}}_{\text{Marginal}} = \frac{1}{p_X(x)} \sum_y y \cdot \underbrace{p_{X,Y}(x,y)}_{\text{Joint}} \end{aligned} \quad (1)$$

(1) 相当于求加权平均数。

从几何角度来看，如图 3 所示，条件概率质量函数 $p_{Y|X}(y|x)$ 分别乘以对应 y 值（绿色高亮），然后求和，结果就是条件期望 $E(Y|X = x)$ 。

图 3. 条件概率 PMF $p_{Y|X}(y|x)$, X 和 Y 均为离散随机变量

解剖条件期望 $E(Y|X = x)$

下面，我们进一步解剖(1)。

给定 $X = x$ 条件下，也就是说离散随机变量 X 固定在 x ，满足这个条件的样本构成了全新的“样本空间”。

$p_{Y|X}(y|x)$ 是给定 $X = x$ 条件下 Y 的概率质量函数，相当于(1)中加权平均数中的权重。

回忆本书第4章，利用贝叶斯定理， $p_X(x) > 0$ ，条件概率质量函数 $p_{Y|X}(y|x)$ 可以通过联合 PMF $p_{X,Y}(x,y)$ 和边缘 PMF $p_X(x)$ 相除得到：

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{\underbrace{p_X(x)}_{\text{Normalize}}} \quad (2)$$

其中，分母中的边缘概率 $p_X(x)$ 起到归一化的效果。

(1) 中大西格玛求和 $\sum_y (\cdot)$ 代表“穷举”一切可能的 y 值，计算“ $y \times$ 条件概率 $p_{Y|X}(y|x)$ ”之和，也就是“ $y \times$ 权重”之和，即加权平均数。

比较期望 $E(Y)$ 、条件期望 $E(Y|X = x)$

对比离散随机变量 Y 的期望 $E(Y)$ 、条件期望 $E(Y|X = x)$ ：

$$\begin{aligned} E(Y) &= \sum_y y \cdot \underbrace{p_Y(y)}_{\text{Weight}} \\ E(Y|X = x) &= \sum_y y \cdot \underbrace{p_{Y|X}(y|x)}_{\text{Weight}} \end{aligned} \quad (3)$$

容易发现，我们不过是把求均值的权重从边缘 PMF $p_Y(y)$ 换成了条件 PMF $p_{Y|X}(y|x)$ 。

\sum_y 都是遍历所有 y 的取值。

作为权重， $p_Y(y)$ 和 $p_{Y|X}(y|x)$ 的求和都为 1，即：

$$\begin{aligned} \sum_y \underbrace{p_Y(y)}_{\text{Marginal}} &= 1 \\ \sum_y \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}} &= 1 \end{aligned} \quad (4)$$

上两式实际上都是本书第 3 章介绍的全概率定理 (law of total probability) 的体现。

⚠ 注意，期望 $E(Y)$ 是一个标量值。而 $E(Y|X=x)$ 在不同的 $X=x$ 条件下结果不同，即 $E(Y|X)$ 代表一组数。也就是说， $E(Y|X)$ 可以看做是个向量。本书前文提过，求期望 $E()$ 运算相当于“归纳”，降维。也就是说 $E(Y|X)$ 中“ Y ”已经被“压缩”成了一个数值，但是 X 还是可变的。

既然 $E(Y|X)$ 代表一组数，我们立刻就会想到 $E(Y|X)$ 肯定也有期望，即均值！

也就是说，笼子里的鸡的平均体重、兔子的平均体重，这两个均值还能再算一个均值，即笼子里所有动物的平均体重。

全期望定理

全期望定理 (law of total expectation)，又叫双重期望定理 (double expectation theorem)、重叠期望定理 (iterated total expectation)，具体指的是：

$$\underset{\text{Expectation}}{E(Y)} = \underset{\text{Conditional expectation}}{E\left[\underset{\text{Conditional expectation}}{E(Y|X)}\right]} = \sum_x \underset{\text{Conditional expectation}}{E(Y|X=x)} \cdot \underset{\text{Marginal}}{p_X(x)} \quad (5)$$

推导过程如下，不要求大家记忆：

$$\begin{aligned} E\left[\underset{\text{Conditional expectation}}{E(Y|X)}\right] &= \sum_x \underset{\text{Conditional expectation}}{E(Y|X=x)} \cdot \underset{\text{Marginal}}{p_X(x)} = \sum_x \left\{ \sum_y y \cdot \underset{\text{Conditional}}{p_{Y|X}(y|x)} \right\} \cdot \underset{\text{Marginal}}{p_X(x)} \\ &= \sum_x \sum_y y \cdot \underset{\text{Conditional}}{p_{Y|X}(y|x)} \cdot \underset{\text{Marginal}}{p_X(x)} = \sum_x \sum_y y \cdot \underset{\text{Joint}}{p_{X,Y}(x,y)} \\ &\stackrel{\text{Use Bayes' Rule}}{=} \sum_x \sum_y y \cdot \underset{\text{Conditional}}{p_{X|Y}(x|y)} \cdot \underset{\text{Marginal}}{p_Y(y)} = \sum_y y \cdot \underset{\text{Marginal}}{p_Y(y)} \cdot \sum_x \underset{\text{Law of total probability}}{p_{X|Y}(x|y)} \\ &= \sum_y y \cdot \underset{\text{Marginal}}{p_Y(y)} = E(Y) \end{aligned} \quad (6)$$

⚠ 注意，以上推导中，二重求和调换变量顺序，这是因为 x 和 y 构成的网格“方方正正”；否则，不能轻易调换求和顺序。这和调换二重积分变量顺序类似。



《数学要素》第 14 章探讨过这个问题，请大家回顾。

白话说全期望定理

其实，全**期望**定理很好理解！

还是用本章前文的例子。前文提到，笼子里的鸡（60%）的平均体重 2 kg，兔子（40%）的平均体重为 4 kg。整个笼子里所有动物的平均体重就是 $2 \times 60\% + 4 \times 40\% = 2.8 \text{ kg}$ 。

前文提过，2 kg、4 kg 都是条件**期望**。

2.8 kg 就是“条件**期望的期望**”。笼子里的鸡占比较高，因此整个笼子里动物的平均体重稍微“偏向”鸡体重的“条件**期望**”。

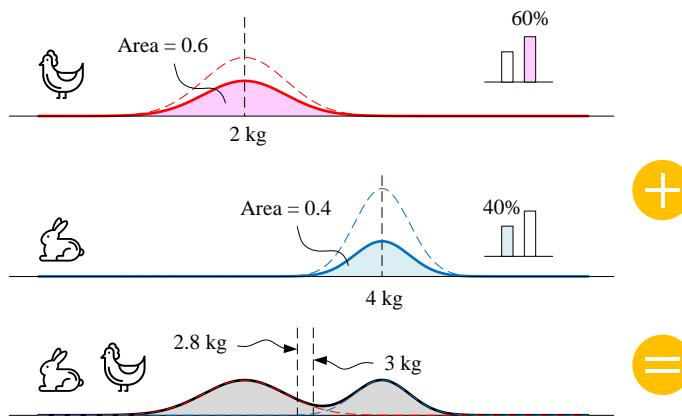


图 4. 解释全**期望**定理

大家如果要问，为什么求“条件**期望的期望**”要用加权平均？而不是用 $(2 + 4) / 2 = 3 \text{ kg}$ ？

为了回答这个问题，我们举个极端例子来解释。除了所有鸡之外，如果整个笼子里只有一只兔子，它的体重为 8 kg，也就是说“所有”兔子的平均体重也是 8 kg。假设所有鸡的平均体重还是 2 kg。大家自己思考，如果用 2 kg 和 8 kg 的平均值 5 kg 代表整个笼子里所有动物的平均体重，这是否合理？

条件**期望** $E(X|Y=y)$

同理，如图 5 所示，给定 $Y=y$ 这个条件下， $p_Y(y) > 0$ ， X 的条件**期望** $E(X|Y=y)$ 定义为：

$$\begin{aligned}
 E(X \mid Y = y) &= \sum_x \underbrace{x \cdot p_{X|Y}(x \mid y)}_{\text{Conditional}} \\
 &= \sum_x \underbrace{x \cdot \frac{p_{X,Y}(x, y)}{p_Y(y)}}_{\text{Marginal}} = \frac{1}{p_Y(y)} \sum_x \underbrace{x \cdot p_{X,Y}(x, y)}_{\text{Joint}}
 \end{aligned} \tag{7}$$

请大家自行分析上式，并比较 $E(X)$ 和 $E(X|Y = y)$ ：

$$\begin{aligned}
 E(X) &= \sum_x x \cdot \underbrace{p_X(x)}_{\text{Weight}} \\
 E(X \mid Y = y) &= \sum_x x \cdot \underbrace{p_{X|Y}(x \mid y)}_{\text{Weight}}
 \end{aligned} \tag{8}$$

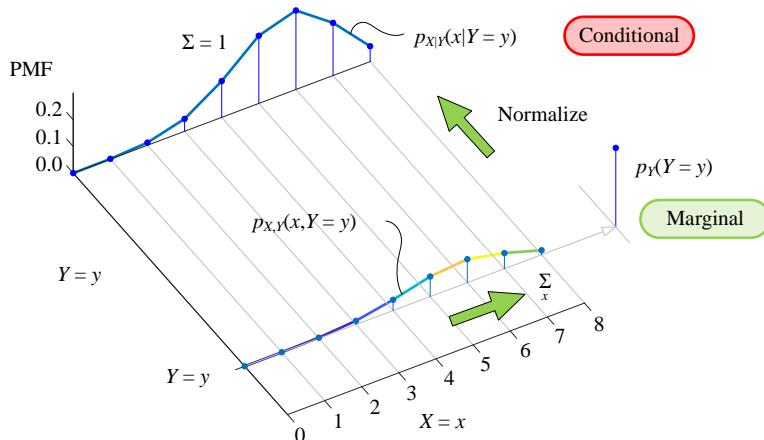


图 5. 条件概率 PMF $p_{X|Y}(x|y)$, X 和 Y 均为离散随机变量

对于条件**期望** $E(X|Y)$, 全**期望**定理为：

$$E(X) = E \left[\underbrace{E(X \mid Y)}_{\text{Conditional expectation}} \right] \tag{9}$$

基于事件的条件期望

给定事件 C 发生的条件下 ($\Pr(C) > 0$), 随机变量 X 的条件**期望**为：

$$\begin{aligned} \mathbb{E}(X|C) &= \sum_x x \cdot \underbrace{p_{X|C}(x|C)}_{\text{Conditional}} \\ &= \sum_x x \cdot \frac{\underbrace{p_{X,C}(x,C)}_{\text{Joint}}}{\Pr(C)} \end{aligned} \quad (10)$$

举个例子，事件 C 可以是鸢尾花数据中指定的标签。

这个式子类似前文的两个随机变量的条件期望，大家会在本章后续看到上式的用途。

独立

特别地，如果 X 和 Y 独立，则：

$$\begin{aligned} \mathbb{E}(Y|X=x) &= \mathbb{E}(Y) \\ \mathbb{E}(X|Y=y) &= \mathbb{E}(X) \end{aligned} \quad (11)$$

8.2 离散随机变量：条件方差

在上一节的基础上，本节介绍离散随机变量的条件方差。

条件方差 $\text{var}(Y|X=x)$

给定 $X=x$ 条件下， Y 的条件方差 $\text{var}(Y|X=x)$ (conditional variance of Y given $X=x$) 定义为：

$$\begin{aligned} \text{var}(Y|X=x) &= \overbrace{\sum_y \left(\underbrace{y - \overbrace{\mathbb{E}(Y|X=x)}^{\text{Expectation}}}_{\text{Deviation}} \right)^2 \cdot \underbrace{p_{Y|X}(y|x)}_{\text{Conditional}}}^{\text{Expectation}} \\ &= \sum_y \left(y - \mathbb{E}(Y|X=x) \right)^2 \cdot \frac{\underbrace{p_{X,Y}(x,y)}_{\text{Joint}}}{\underbrace{p_X(x)}_{\text{Marginal}}} \\ &= \underbrace{\frac{1}{p_X(x)}}_{\text{Marginal}} \sum_y \left(\underbrace{y - \mathbb{E}(Y|X=x)}_{\text{Deviation}} \right)^2 \cdot \underbrace{p_{X,Y}(x,y)}_{\text{Joint}} \end{aligned} \quad (12)$$

下面解剖上式。

$\mathbb{E}(Y|X=x)$ 是 (1) 中求得的条件期望，也就是计算偏差的基准。

$y - E(Y|X=x)$ 代表偏差，即每个 y 和 $E(Y|X=x)$ 之间的偏离。 $y - E(Y|X=x)$ 平方后，再以 $p_{Y|X}(y|x)$ 为权重，求平均值，结果就是条件**方差**。

对比离散随机变量 Y 的**方差**和条件**方差**：

$$\begin{aligned} \text{var}(Y) &= \sum_y \left(\underbrace{y - E(Y)}_{\text{Deviation}} \right)^2 \cdot \underbrace{p_Y(y)}_{\text{Weight}} \\ \text{var}(Y) &= \sum_y \left(\underbrace{y - E(Y|X=x)}_{\text{Deviation}} \right)^2 \cdot \underbrace{p_{Y|X}(y|x)}_{\text{Weight}} \end{aligned} \quad (13)$$

可以发现两处变差异，度量偏差的基准从 $E(Y)$ 变成 $E(Y|X=x)$ 。加权平均的权重从 $p_Y(y)$ 变成 $p_{Y|X}(y|x)$ 。

类似**方差**的简便计算技巧，条件**方差** $\text{var}(Y|X=x)$ 也有如下计算技巧：

$$\begin{aligned} \text{var}(Y) &= E(Y^2) - E(Y)^2 \\ \text{var}(Y|X=x) &= E(Y^2|X=x) - E(Y|X=x)^2 \end{aligned} \quad (14)$$

全**方差定理**

全方差定理 (law of total variance)，又叫**重叠期望定理** (law of iterated variance)，指的是：

$$\text{var}(Y) = \underbrace{E(\text{var}(Y|X))}_{\text{Expectation of conditional variance}} + \underbrace{\text{var}(E(Y|X))}_{\text{Variance of conditional expectation}} \quad (15)$$

$E(\text{var}(Y|X))$ 是条件**方差**的**期望** (加权平均数)：

$$\underbrace{E(\text{var}(Y|X))}_{\text{Expectation of conditional variance}} = \sum_x \text{var}(Y|X=x) \cdot p_X(x) \quad (16)$$

条件**方差**的**期望** $E(\text{var}(Y|X))$ 还不够解释整体的**方差**。缺少的成分是条件**期望的方差** $\text{var}(E(Y|X))$ ：

$$\underbrace{\text{var}(E(Y|X))}_{\text{Variance of conditional expectation}} = \sum_x (E(Y|X=x) - E(Y))^2 \cdot p_X(x) \quad (17)$$

根据全**期望定理**， $E(Y|X=x)$ 的**期望**为 $E(Y)$ 。

换个方向思考，(15) 相当于对 $\text{var}(Y)$ 的分解：

$$\begin{aligned}\text{var}(Y) &= \underbrace{\mathbb{E}(\text{var}(Y|X))}_{\text{Expectation of conditional variance}} + \underbrace{\text{var}(\mathbb{E}(Y|X))}_{\text{Variance of conditional expectation}} \\ &= \sum_x \underbrace{\text{var}(Y|X=x)}_{\text{Deviation within a subset}} \cdot p_X(x) + \sum_x \underbrace{\left(\overbrace{\mathbb{E}(Y|X=x) - \mathbb{E}(Y)}^{\text{Deviation of a subset from superset}} \right)^2 \cdot p_X(x)}_{\text{Deviation among all subsets}} \quad (18)\end{aligned}$$

这方便我们理解哪些成分（子集内部、子集之间）以多大的比例贡献了整体的**方差**。

如图 6 所示，条件**方差的期望**解释的是子集（鸡子集、兔子集）各自内部差异。

条件**期望的方差**解释的是子集（鸡子集、兔子集）和母集（所有动物）之间的差异。

而代表鸡子集、兔子集就是鸡、兔各自的平均体重（条件**期望**），代表母集就是笼子里所有动物的平均体重（总体**期望**）。

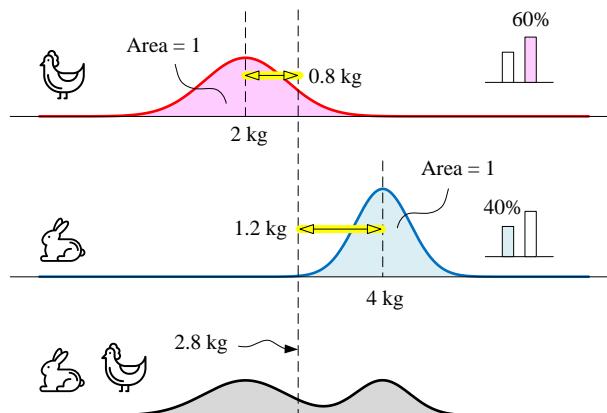


图 6. 解释全**方差**定理

比较图 6 和图 7，条件**方差的期望**不变，但是条件**期望的方差**增大。如图 7 所示，子集内部差异（**方差**）不变，如果增大子集之间的差异，也就是增大了子集和母集的差异，这会导致整体的**方差**增大。

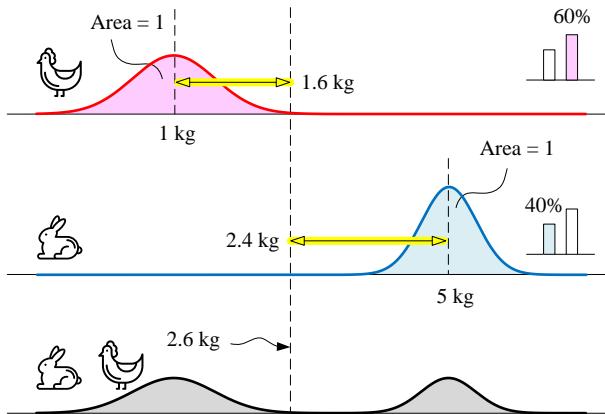


图 7. 解释全方差定理，增大子集之间差异，整体方差增大

类似全方差定理，也存在如下**全协方差定理** (law of total covariance):

$$\text{cov}(X_1, X_2) = E(\text{cov}(X_1, X_2 | Y)) + \text{cov}(E(X_1 | Y), E(X_2 | Y)) \quad (19)$$

本章不展开分析全协方差定理。

条件方差 $\text{var}(X|Y=y)$

给定 $Y=y$ 条件下， X 的条件方差 $E(X|Y=y)$ (conditional variance of X given $Y=y$) 定义为：

$$\begin{aligned} \text{var}(X|Y=y) &= \sum_x \underbrace{\left(\underbrace{x - \overbrace{E(X|Y=y)}^{\text{Expectation}}}_{\text{Deviation}} \right)^2}_{\text{Conditional}} \cdot p_{X|Y}(x|y) \\ &= \sum_x \left(x - E(X|Y=y) \right)^2 \cdot \frac{\overbrace{p_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{p_Y(y)}_{\text{Marginal}}} \\ &= \underbrace{\frac{1}{p_Y(y)}}_{\text{Marginal}} \sum_x \underbrace{\left(x - E(X|Y=y) \right)^2}_{\text{Deviation}} \cdot \overbrace{p_{X,Y}(x,y)}^{\text{Joint}} \end{aligned} \quad (20)$$

条件方差 $\text{var}(X|Y=y)$ 也有如下计算技巧：

$$\text{var}(X|Y=y) = E(X^2|Y=y) - E(X|Y=y)^2 \quad (21)$$

对于随机变量 X ，它的全方差定理为：

$$\text{var}(X) = \underbrace{E(\text{var}(X|Y))}_{\text{Expectation of conditional variance}} + \underbrace{\text{var}(E(X|Y))}_{\text{Variance of conditional expectation}} \quad (22)$$

8.3 离散随机变量条件期望、条件方差：以鸢尾花为例

给定花萼长度，条件期望 $E(X_2 | X_1 = x_1)$

大家已经在本书第4章见过图8中左图。这幅图给出的是条件概率 $p_{X_2|X_1}(x_2 | x_1)$ 。提醒大家回忆，图中 $p_{X_2|X_1}(x_2 | x_1)$ 每列 PMF(即概率) 和为 1，即满足(4)。

下面，我们试着利用图8中左图计算花萼长度 $X_1 = 6.5$ 为条件下，条件期望 $E(X_2 | X_1 = 6.5)$ ：

$$\begin{aligned} E(X_2 | X_1 = 6.5) &= \sum_{x_2} x_2 \cdot p_{X_2|X_1}(x_2 | 6.5) \\ &= 2.0 \times 0 + 2.5 \times 0.19 + 3.0 \times 0.65 + 3.5 \times 0.16 + 4.0 \times 0 + 4.5 \times 0 \\ &\approx 2.984 \text{ cm} \end{aligned} \quad (23)$$

注意，上式中条件概率的结果还是 cm。

建议大家手算剩余所有 $E(X_2 | X_1 = x_1)$ 。

图8中右上图给出的是热图 $x_2 \cdot p_{X_2|X_1}(x_2 | x_1)$ ，相当于一个二元函数。

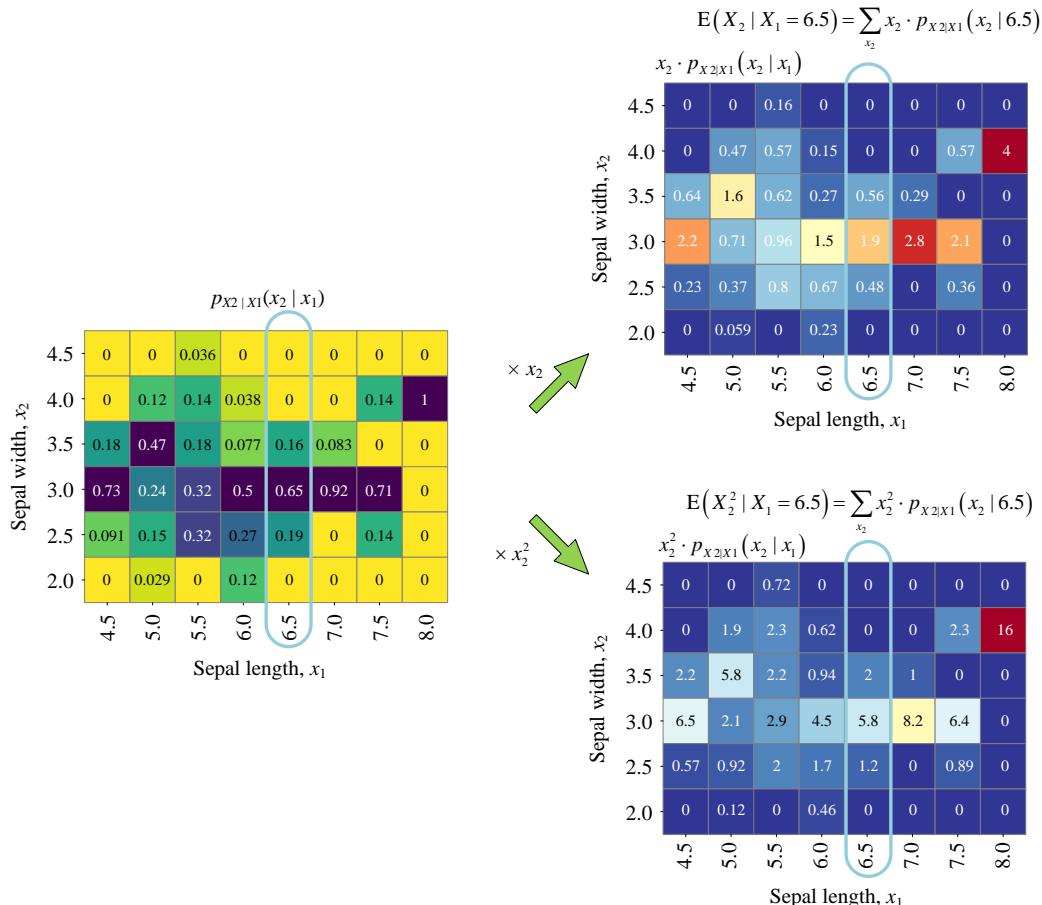
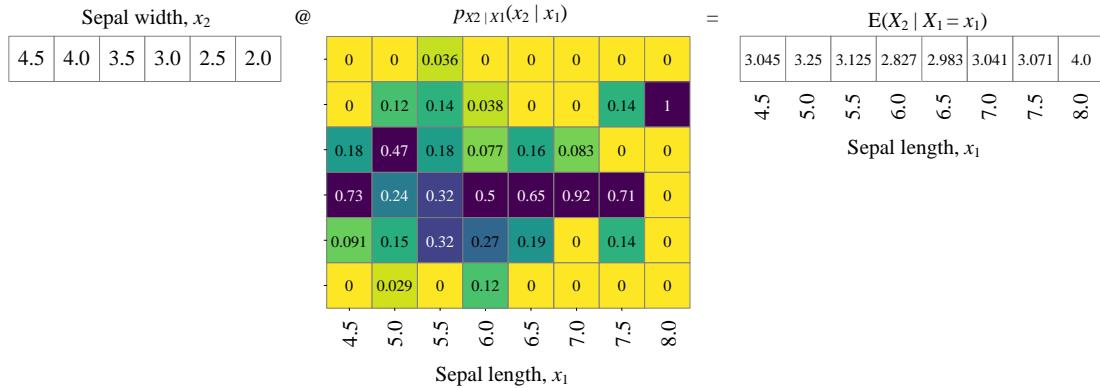
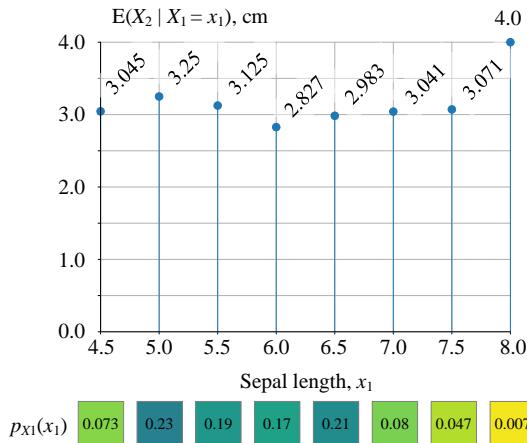


图 8. 给定花萼长度 X_1 , 花萼宽度 X_2 的条件概率 $p_{X_2|X_1}(x_2|x_1)$ 热图, $x_2 \times p_{X_2|X_1}(x_2|x_1)$ 热图, $x_2^2 \cdot p_{X_2|X_1}(x_2|x_1)$ 热图

图 9 所示为从矩阵乘法视角看条件期望 $E(X_2 | X_1 = x_1)$ 运算。

图 10 所示为条件期望 $E(X_2 | X_1 = x_1)$ 的火柴梗图。图 10 中还给出了鸢尾花花萼长度 X_1 的边缘 PMF $p_{X_1}(x_1)$ 。

图 9. 矩阵乘法视角看条件期望 $E(X_2 | X_1 = x_1)$ 图 10. 给定花萼长度 X_1 , 花萼宽度 X_2 的条件期望 $E(X_2 | X_1 = x_1)$, 和边缘 PMF $p_{X_1}(x_1)$

根据(5)的全期望定理, 我们可以利用条件期望 $E(X_2 | X_1 = x_1)$ 和边缘 PMF $p_{X_1}(x_1)$ 计算期望 $E(X_2)$:

$$\begin{aligned} E(X_2) &= \sum_{x_1} E(X_2 | X_1 = x_1) \cdot p_{X_1}(x_1) \\ &= 3.045 \times 0.073 + 3.25 \times 0.23 + 3.125 \times 0.19 + 2.827 \times 0.17 + \\ &\quad 2.983 \times 0.21 + 3.041 \times 0.08 + 3.071 \times 0.047 + 4 \times 0.007 \\ &\approx 3.063 \text{ cm} \end{aligned} \tag{24}$$

给定花萼长度，条件方差 $\text{var}(X_2 | X_1 = x_1)$

利用(12)计算花萼长度 $X_1 = 6.5$ 为条件下，条件方差 $\text{var}(X_2 | X_1 = 6.5)$ ：

$$\begin{aligned}\text{var}(X_2 | X_1 = 6.5) &= \sum_{x_2} (x_2 - E(X_2 | X_1 = 6.5))^2 \cdot p_{X_2|X_1}(x_2 | 6.5) \\ &= \underbrace{(2.0 - 2.985)^2}_{\text{cm}^2} \times 0 + \underbrace{(2.5 - 2.985)^2}_{\text{cm}^2} \times 0.19 + \underbrace{(3.0 - 2.985)^2}_{\text{cm}^2} \times 0.65 + \\ &\quad \underbrace{(3.5 - 2.985)^2}_{\text{cm}^2} \times 0.16 + \underbrace{(4.0 - 2.985)^2}_{\text{cm}^2} \times 0 + \underbrace{(4.5 - 2.985)^2}_{\text{cm}^2} \times 0 \\ &\approx 0.088 \text{ cm}^2\end{aligned}\tag{25}$$

条件方差 $\text{var}(X_2 | X_1 = 6.5)$ 的单位为 cm^2 。同样建议大家手算剩余条件方差 $\text{var}(X_2 | X_1 = x_1)$ 。

采用技巧计算，计算条件期望。首先计算花萼长度 $X_1 = 6.5$ 为条件下，花萼宽度平方的期望：

$$\begin{aligned}E(X_2^2 | X_1 = 6.5) &= \sum_{x_2} x_2^2 \cdot p_{X_2|X_1}(x_2 | 6.5) \\ &= \underbrace{2.0^2}_{\text{cm}^2} \times 0 + \underbrace{2.5^2}_{\text{cm}^2} \times 0.19 + \underbrace{3.0^2}_{\text{cm}^2} \times 0.65 + \underbrace{3.5^2}_{\text{cm}^2} \times 0.16 + \underbrace{4.0^2}_{\text{cm}^2} \times 0 + \underbrace{4.5^2}_{\text{cm}^2} \times 0 \\ &\approx 9 \text{ cm}^2\end{aligned}\tag{26}$$

图 11 所示为花萼宽度平方值 X_2^2 的条件期望 $E(X_2^2 | X_1 = x_1)$ 的火柴梗图。

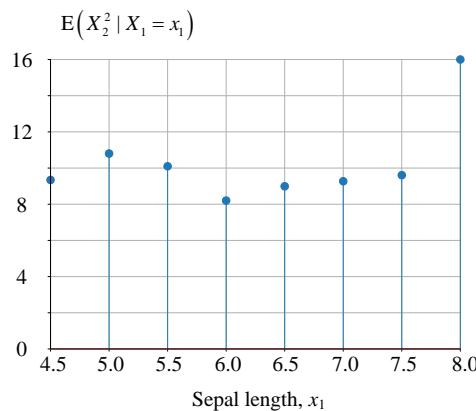


图 11. 给定花萼长度 X_1 ，花萼宽度平方值 X_2^2 的条件期望 $E(X_2^2 | X_1 = x_1)$

然后计算条件方差：

$$\text{var}(X_2 | X_1 = 6.5) = E(X_2^2 | X_1 = 6.5) - E(X_2 | X_1 = 6.5)^2 = 9 - 2.984^2 \approx 0.088\tag{27}$$

图 12 所示为花萼长度取不同值时条件方差 $\text{var}(X_2 | X_1 = x_1)$ 的火柴梗图，请大家用作检查自己手算结果。

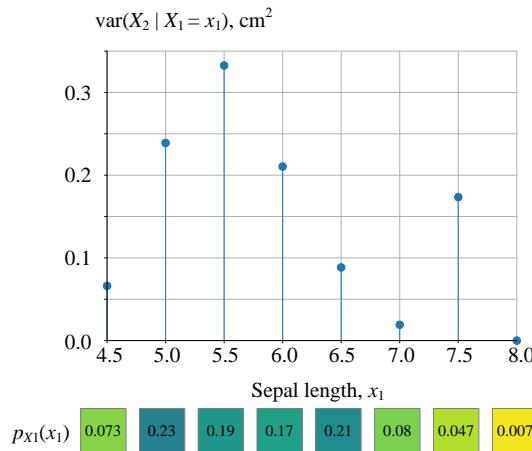


图 12. 给定花萼长度 X_1 , 花萼宽度的条件方差 $\text{var}(X_2 | X_1 = x_1)$

大家肯定早就发现，条件期望 $E(X_2 | X_1 = x_1)$ 、条件方差 $\text{var}(X_2 | X_1 = x_1)$ 都消去了 x_2 这个变量，两者仅仅随着 $X_1 = x_1$ 取值变化。这也不难理解，**期望**和**方差**代表“汇总”，本质上就是“降维”。某个维度上的信息细节不再重要，我们把这个“压扁”。

压扁过程中，不同的聚合方式得到不同的统计量，比如**期望**、**方差**等等。

全方差定理：还原方差 $\text{var}(X_2)$

根据 (17) 中给出的全**方差**定理，下面我们利用条件**方差** $\text{var}(X_2 | X_1)$ 和条件**期望** $E(X_2 | X_1)$ 计算花萼宽度的**方差** $\text{var}(X_2)$ 。 $\text{var}(X_2)$ 可以写成两部分之和：

$$\text{var}(X_2) = \underbrace{\mathbb{E}(\text{var}(X_2 | X_1))}_{\text{Expectation of conditional variance}} + \underbrace{\text{var}(\mathbb{E}(X_2 | X_1))}_{\text{Variance of conditional expectation}} \quad (28)$$

第一部分是条件**方差**的**期望** $E(\text{var}(X_2 | X_1))$ ：

$$\underbrace{\mathbb{E}(\text{var}(X_2 | X_1))}_{\text{Expectation of conditional variance}} = \sum_{x_1} \text{var}(X_2 | X_1 = x_1) \cdot p_{x1}(x_1) \quad (29)$$

代入具体数值，我们可以计算得到 $E(\text{var}(X_2 | X_1))$ ：

$$\begin{aligned}
 \underbrace{\mathbb{E}(\text{var}(X_2 | X_1))}_{\text{Expectation of conditional variance}} &= \sum_{x_1} \text{var}(X_2 | X_1 = x_1) \cdot p_{X_1}(x_1) \\
 &\approx 0.066 \times 0.073 + 0.238 \times 0.226 + 0.332 \times 0.186 + 0.210 \times 0.173 + \\
 &\quad 0.088 \times 0.206 + 0.019 \times 0.08 + 0.173 \times 0.046 + 0 \times 0.006 \\
 &\approx \underbrace{0.0048}_{X_1=4.5} + \underbrace{0.0541}_{X_1=5.0} + \underbrace{0.0620}_{X_1=5.5} + \underbrace{0.0364}_{X_1=6.0} + \underbrace{0.0182}_{X_1=6.5} + \underbrace{0.0015}_{X_1=7.0} + \underbrace{0.0080}_{X_1=7.5} + \underbrace{0}_{X_1=8.0} \\
 &\approx 0.185 \text{ cm}^2
 \end{aligned} \tag{30}$$

第二部分是条件期望的方差 $\text{var}(\mathbb{E}(X_2 | X_1))$ 。代入具体值计算得到：

$$\begin{aligned}
 \underbrace{\text{var}(\mathbb{E}(X_2 | X_1))}_{\text{Variance of conditional expectation}} &= \sum_{x_1} (\mathbb{E}(X_2 | X_1 = x_1) - \mathbb{E}(X_2))^2 \cdot p_{X_1}(x_1) \\
 &\approx 0.025 \text{ cm}^2
 \end{aligned} \tag{31}$$

如果大家看到这还会犯糊涂，不理解为什么 \sum_{x_1} 求和遍历的是 x_1 ？

告诉大家一个小技巧，因为 X_2 已经被“折叠”！不管是条件期望 $\mathbb{E}(X_2 | X = x_1)$ 、还是期望 $\mathbb{E}(X_2)$ ，都已经将 X_2 折叠成一个具体的数值，因此无法遍历。

这样 X_2 的方差约为：

$$\begin{aligned}
 \text{var}(X_2) &= \underbrace{\mathbb{E}(\text{var}(X_2 | X_1))}_{\text{Expectation of conditional variance}} + \underbrace{\text{var}(\mathbb{E}(X_2 | X_1))}_{\text{Variance of conditional expectation}} \\
 &\approx 0.185 + 0.025 = 0.211 \text{ cm}^2
 \end{aligned} \tag{32}$$

在 $\text{var}(X_2)$ 中，第一部分 $\mathbb{E}(\text{var}(X_2 | X_1))$ 贡献超过 85%。而 $\text{var}(\mathbb{E}(X_2 | X_1))$ 可以进一步展开，图 13 所示为各个不同成分对花萼宽度 X_2 的方差 $\text{var}(X_2)$ 的贡献，这也叫做钻取 (drill down)。

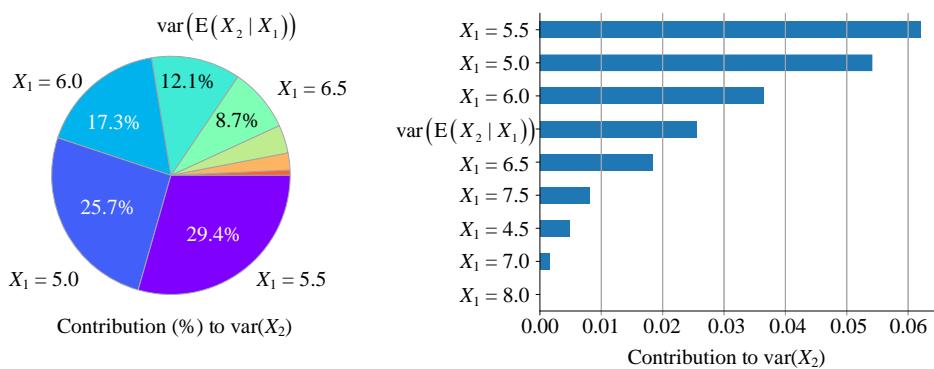


图 13. 各个不同成分对花萼宽度 X_2 的方差 $\text{var}(X_2)$ 的贡献

给定花萼长度，条件标准差 $\text{std}(X_2 | X_1 = x_1)$

(25) 开方便获得条件**标准差** $\text{std}(X_2 | X_1 = 6.5)$:

$$\sigma_{X_2|X_1=6.5} = \text{std}(X_2 | X_1 = 6.5) = 0.295 \text{ cm} \quad (33)$$

上式的单位和鸢尾花宽度单位一致，我们便可以把条件**标准差**和图 10 画在一起，得到图 14。这幅图给出的是 $E(X_2 | X_1 = x_1) \pm \text{std}(X_2 | X_1 = x_1)$ 。

圆点 ● 展示的是 $E(X_2 | X_1 = x_1)$ ，即条件**期望**，代表给定 $X_1 = x_1$ 条件下，鸢尾花数据在花萼宽度上的一种“预测”！这和我们讲过的回归思想本质上相同。 $E(X_2 | X_1 = x_1)$ 代表当 $X_1 = x_1$ 时鸢尾花花萼宽度最合适的“预测”。

也就是说，回归可以看成是条件概率！本书后续还会沿着这个思路展开讨论。

而我们用**误差棒** (error bar) 展示 $\pm \text{std}(X_2 | X_1 = x_1)$ ，代表给定 $X_1 = x_1$ 条件下，鸢尾花数据在花萼宽上的“波动”。误差棒的宽度越大，说明波动越大；反之，则说明波动越小。

特别地，当花萼长度 X_1 为 8.0 cm 时，条件均**方差** $\text{std}(X_2 | X_1 = 8.0)$ 为 0。这是因为，这一处只有一个样本点。

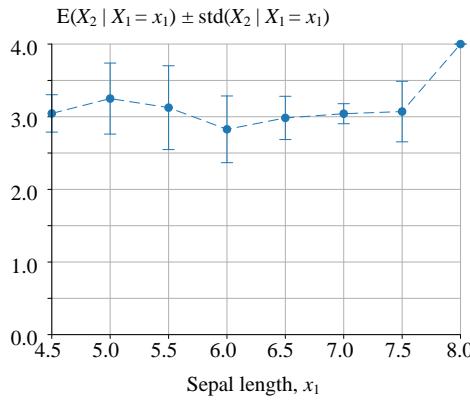


图 14. 给定花萼长度 X_1 ，花萼宽度 X_2 的条件**期望** $E(X_2 | X_1 = x_1) \pm \text{std}(X_2 | X_1 = x_1)$

给定花萼宽度，条件**期望** $E(X_1 | X_2 = x_2)$

图 15 给出的是条件概率 $p_{X1/X2}(x_1 | x_2)$ 。同样提醒大家注意图中 $p_{X1/X2}(x_1 | x_2)$ 每行 PMF (即概率) 和为 1。

利用图 15 计算花萼宽度 $X_2 = 2.0$ 为条件下，条件**期望** $E(X_1 | X_2 = 2.0)$:

$$\begin{aligned} E(X_1 | X_2 = 2.0) &= \sum_{x_1} x_1 \cdot p_{X1|X2}(x_1 | 2.0) \\ &= 4.5 \times 0 + 5.0 \times 0.25 + 5.5 \times 0 + 6.0 \times 0.75 + \\ &\quad 6.5 \times 0 + 7.0 \times 0 + 7.5 \times 0 + 8.0 \times 0 \\ &\approx 5.7 \text{ cm} \end{aligned} \quad (34)$$

条件概率的结果还是 cm。同样建议大家手算剩余所有 $E(X_1 | X_2 = x_2)$ 。

此外，请大家也根据全**期望**定理，利用 $E(X_1 | X_2 = x_2)$ 计算 $E(X_1)$ 。并用条件**方差** $\text{var}(X_1 | X_2)$ 和条件**期望** $E(X_1 | X_2)$ 计算花萼长度的**方差** $\text{var}(X_1)$ 。

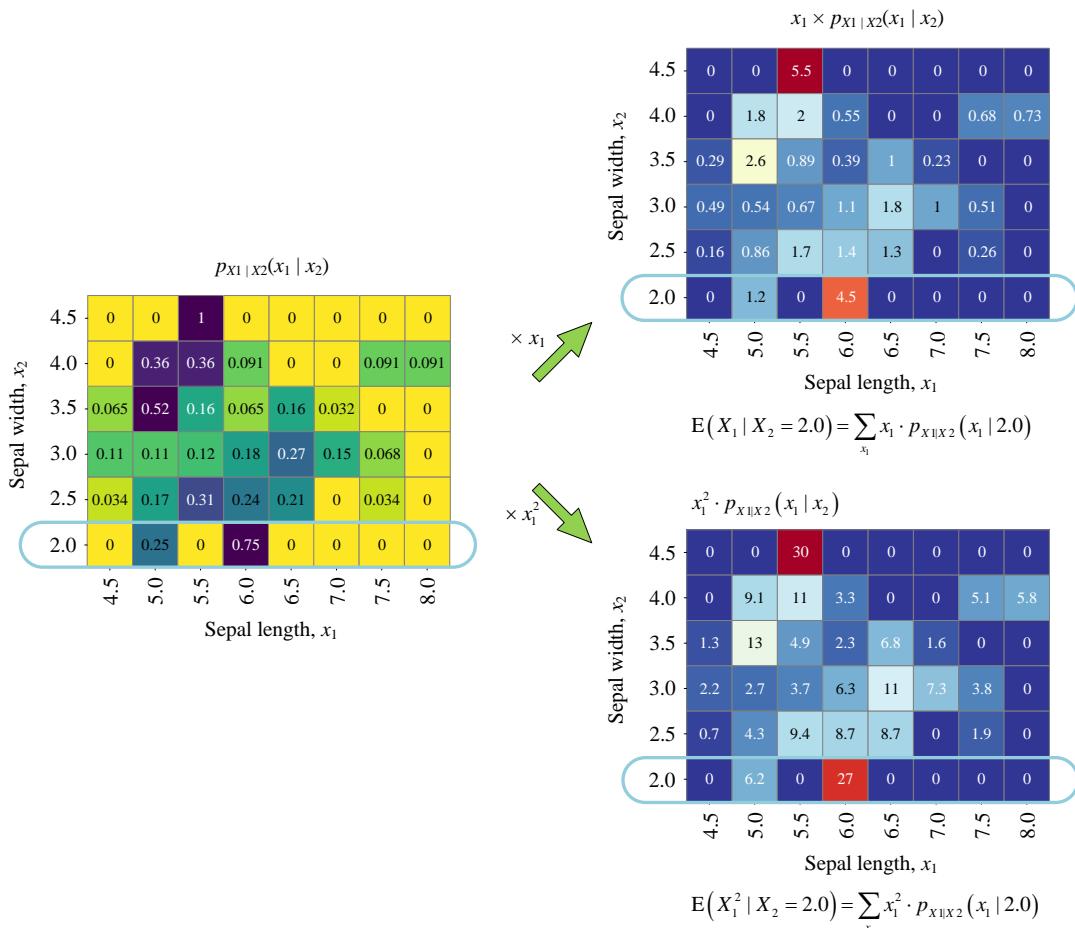


图 15. 给定花萼宽度，花萼长度的条件概率 $p_{X_1|X_2}(x_1 | x_2)$

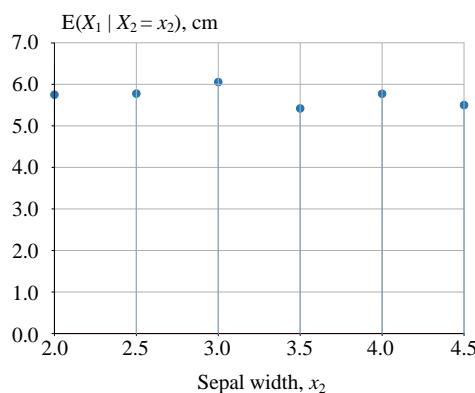


图 16. 给定花萼宽度 X_2 , 花萼宽度的条件期望 $E(X_1 | X_2 = x_2)$

条件方差 $\text{var}(X_1 | X_2 = x_2)$

在花萼宽度 $X_2 = 2.0$ 为条件下, 条件方差 $\text{var}(X_1 | X_2 = 2.0)$:

$$\begin{aligned} \text{var}(X_1 | X_2 = 2.0) &= \sum_{x_1} (x_1 - E(X_1 | X_2 = 2.0))^2 \cdot p_{X_1|X_2}(x_1 | 2.0) \\ &= \underbrace{(4.5 - 5.75)^2}_{\text{cm}^2} \times 0 + \underbrace{(5.0 - 5.75)^2}_{\text{cm}^2} \times 0.25 + \underbrace{(5.5 - 5.75)^2}_{\text{cm}^2} \times 0 + \underbrace{(6.0 - 5.75)^2}_{\text{cm}^2} \times 0.75 + \\ &\quad \underbrace{(6.5 - 5.75)^2}_{\text{cm}^2} \times 0 + \underbrace{(7.0 - 5.75)^2}_{\text{cm}^2} \times 0 + \underbrace{(7.5 - 5.75)^2}_{\text{cm}^2} \times 0 + \underbrace{(8.0 - 5.75)^2}_{\text{cm}^2} \times 0 \\ &= 0.1875 \text{ cm}^2 \end{aligned} \quad (35)$$

条件方差 $\text{var}(X_1 | X_2 = 2.0)$ 的单位为 cm^2 。同样建议大家手算剩余条件方差 $\text{var}(X_1 | X_2 = x_2)$ 。

利用条件方差计算技巧, 首先计算花萼宽度 $X_2 = 2.0$ 为条件下, 花萼长度平方的期望:

$$\begin{aligned} E(X_1^2 | X_2 = 2.0) &= \sum_{x_1} x_1^2 \cdot p_{X_1|X_2}(x_1 | 2.0) \\ &= \underbrace{4.5^2}_{\text{cm}^2} \times 0 + \underbrace{5.0^2}_{\text{cm}^2} \times 0.25 + \underbrace{5.5^2}_{\text{cm}^2} \times 0 + \underbrace{6.0^2}_{\text{cm}^2} \times 0.75 + \\ &\quad \underbrace{6.5^2}_{\text{cm}^2} \times 0 + \underbrace{7.0^2}_{\text{cm}} \times 0 + \underbrace{7.5^2}_{\text{cm}^2} \times 0 + \underbrace{8.0^2}_{\text{cm}^2} \times 0 \\ &= 33.25 \text{ cm}^2 \end{aligned} \quad (36)$$

图 17 所示为给定花萼长度 X_2 , 花萼宽度平方值 X_1^2 的条件期望 $E(X_1^2 | X_2 = x_2)$ 。请大家自行代入计算条件方差 $\text{var}(X_1 | X_2 = 2.0)$ 。

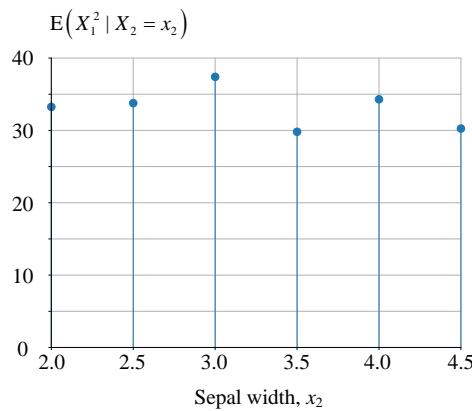
图 17. 给定花萼长度 X_2 , 花萼宽度平方值 X_1^2 的条件期望 $E(X_1^2 | X_2 = x_2)$

图 18 所示为条件方差 $\text{var}(X_1 | X_2 = x_2)$ 的火柴梗图。同样，条件期望 $E(X_1 | X_2 = x_2)$ 、条件方差 $\text{var}(X_1 | X_2 = x_2)$ 都“折叠”了 x_1 这个维度，两者仅仅随着 $X_2 = x_2$ 取值变化。

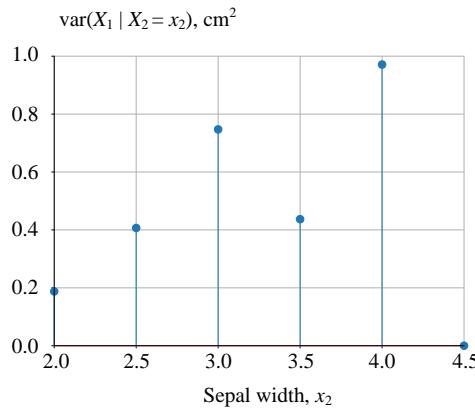


图 18. 给定花萼宽度 X_2 , 花萼宽度的条件方差 $\text{var}(X_1 | X_2 = x_2)$

类似图 14, 我们也绘制给定花萼宽度 X_2 , 花萼长度 X_1 的条件期望 $E(X_1 | X_2 = x_2) \pm \text{std}(X_1 | X_2 = x_2)$ 。请大家自行分析这幅图像。

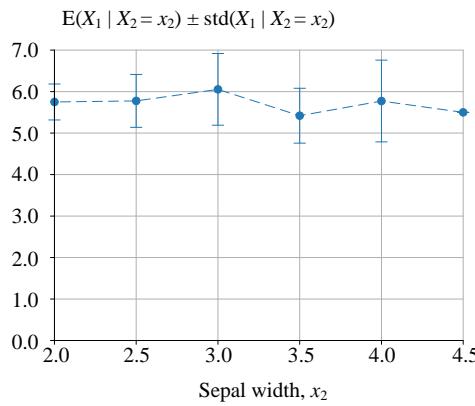


图 19. 给定花萼宽度 X_2 , 花萼长度 X_1 的条件期望 $E(X_1 | X_2 = x_2) \pm \text{std}(X_1 | X_2 = x_2)$

考虑标签：花萼长度

给定鸢尾花分类标签 $Y = C_1$, 花萼长度 X_1 的条件期望:

$$\begin{aligned}
 E(X_1 | Y = C_1) &= \sum_{x_1} x_1 \cdot p_{X_1|Y}(x_1 | C_1) \\
 &= 4.5 \times 0.22 + 5.0 \times 0.56 + 5.5 \times 0.2 + 6.0 \times 0.02 + \\
 &\quad 6.5 \times 0 + 7.0 \times 0 + 7.5 \times 0 + 8.0 \times 0 \\
 &= 5.01 \text{ cm}
 \end{aligned} \tag{37}$$

给定鸢尾花分类标签 $Y = C_1$, 花萼长度 X_1 平方期望:

$$\begin{aligned}
 E(X_1^2 | Y = C_1) &= \sum_{x_1} x_1^2 \cdot p_{X_1|Y}(x_1 | C_1) \\
 &= 4.5^2 \times 0.22 + 5.0^2 \times 0.56 + 5.5^2 \times 0.2 + 6.0^2 \times 0.02 + \\
 &\quad 6.5^2 \times 0 + 7.0^2 \times 0 + 7.5^2 \times 0 + 8.0^2 \times 0 \\
 &= 25.225 \text{ cm}^2
 \end{aligned} \tag{38}$$

给定鸢尾花分类标签 $Y = C_1$, 花萼长度 X_1 条件方差:

$$\begin{aligned}
 \text{var}(X_1 | Y = C_1) &= E(X_1^2 | Y = C_1) - E(X_1 | Y = C_1)^2 \\
 &= 25.225 - 5.01^2 \\
 &= 0.1249 \text{ cm}^2
 \end{aligned} \tag{39}$$

给定鸢尾花分类标签 $Y = C_1$, 花萼长度 X_1 条件标准差:

$$\sigma_{X_1|Y=C_1} = \sqrt{\text{var}(X_1 | Y = C_1)} = \sqrt{0.1249} = 0.353 \text{ cm} \tag{40}$$

请大家自行计算剩余两种情况 ($Y = C_2, C_3$)。并利用全期望定理, 计算 $E(X_1)$ 。

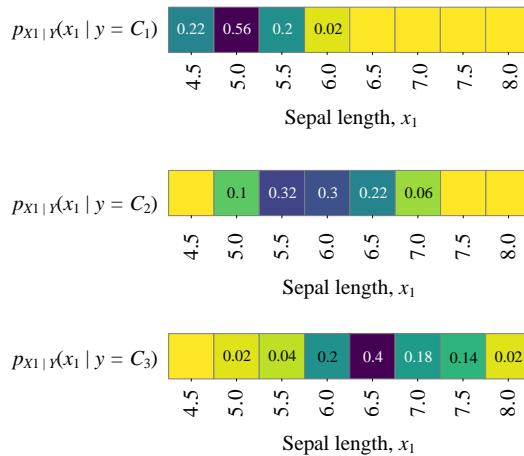


图 20. 给定鸢尾花标签 Y , 花萼长度的条件 PMF

考虑标签：花萼宽度

给定鸢尾花分类标签 $Y = C_1$, 花萼宽度 X_2 的条件**期望**:

$$\begin{aligned} E(X_2 | Y = C_1) &= \sum_{x_2} x_2 \cdot p_{X_2|Y}(x_2 | C_1) \\ &= 4.5 \times 0.07 + 4.0 \times 0.18 + 3.5 \times 0.46 + 3.0 \times 0.32 + 2.5 \times 0.02 + 2.0 \times 0 \\ &= 3.43 \text{ cm} \end{aligned} \quad (41)$$

给定鸢尾花分类标签 $Y = C_1$, 花萼宽度 X_2 平方**期望**:

$$\begin{aligned} E(X_2^2 | Y = C_1) &= \sum_{x_2} x_2^2 \cdot p_{X_2|Y}(x_2 | C_1) \\ &= 4.5^2 \times 0.07 + 4.0^2 \times 0.18 + 3.5^2 \times 0.46 + 3.0^2 \times 0.32 + 2.5^2 \times 0.02 + 2.0^2 \times 0 \\ &= 11.925 \text{ cm}^2 \end{aligned} \quad (42)$$

给定鸢尾花分类标签 $Y = C_1$, 花萼宽度 X_2 条件**方差**:

$$\begin{aligned} \text{var}(X_2 | Y = C_1) &= E(X_2^2 | Y = C_1) - E(X_2 | Y = C_1)^2 \\ &= 11.925 - 3.43^2 \\ &= 0.1601 \text{ cm}^2 \end{aligned} \quad (43)$$

给定鸢尾花分类标签 $Y = C_1$, 花萼宽度 X_2 条件**标准差**:

$$\sigma_{X_2|Y=C_1} = \sqrt{\text{var}(X_2 | Y = C_1)} = \sqrt{0.1601} \approx 0.4 \text{ cm} \quad (44)$$

请大家自行计算鸢尾花其他标签条件下花萼长度、花萼宽度的条件**期望**、条件**方差**、条件**标准差**。

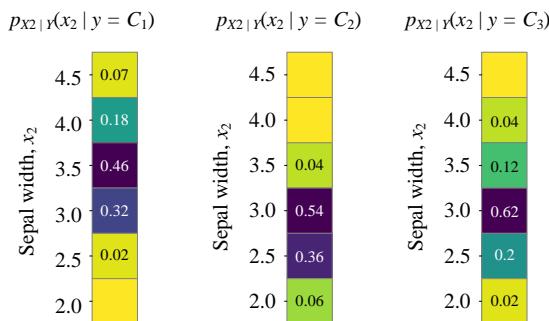


图 21. 给定鸢尾花标签 Y , 花萼宽度的条件 PMF



Bk5_Ch08_01.py 代码绘制本节大部分图像。代码中用到了矩阵乘法和广播原则, 请大家注意区分。

8.4 连续随机变量：条件期望

本节介绍如何计算连续随机变量的条件期望。

条件期望 $E(Y|X = x)$

如果 X 和 Y 均为连续随机变量，如图 22 所示，在给定 $X = x$ 条件下，**条件期望** $E(Y|X = x)$ 定义为：

$$\begin{aligned} E(Y|X = x) &= \overbrace{\int_{-\infty}^{+\infty} y \cdot \underbrace{f_{Y|X}(y|x)}_{\text{Conditional}} dy}^{\text{Expectation}} \\ &= \int_{-\infty}^{+\infty} y \cdot \underbrace{\frac{f_{X,Y}(x,y)}{f_X(x)}}_{\substack{\text{Joint} \\ \text{Marginal}}} dy = \underbrace{\frac{1}{f_X(x)}}_{\text{Marginal}} \int_{-\infty}^{+\infty} y \cdot \underbrace{f_{X,Y}(x,y)}_{\text{Joint}} dy \end{aligned} \quad (45)$$

上式中，边缘概率 $f_X(x)$ 可以通过下式得到：

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy \quad (46)$$

(46) 代入 (45) 得到：

$$E(Y|X = x) = \frac{1}{\int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy} \int_{-\infty}^{+\infty} y \cdot f_{X,Y}(x,y) dy \quad (47)$$

上式，相当于消去了 y ，这和本章前文提到的“降维”、“折叠”本质上没有任何区别。对于离散随机变量，折叠用的数学工具为求和符号 Σ ；连续随机变量则用积分符号 \int 。

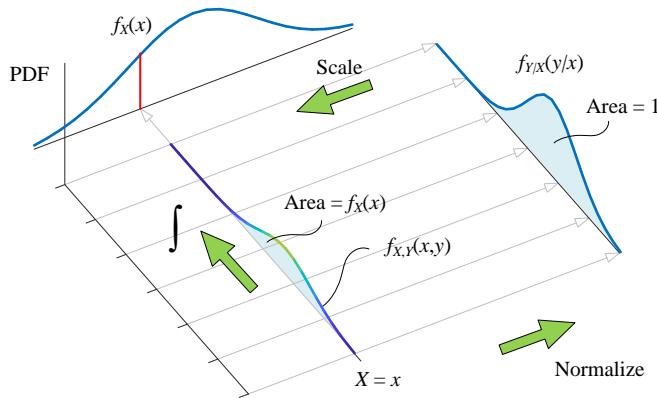


图 22. 联合概率 PDF $f_{X,Y}(x,y)$ 和条件概率 PDF $f_{Y|X}(y|x)$ 的关系， X 和 Y 均为连续随机变量

条件期望 $E(X|Y=y)$

同理，如图 23 所示，条件期望 $E(X|Y=y)$ 定义为：

$$E(X|Y=y) = \frac{1}{\int_{-\infty}^{+\infty} f_{X,Y}(x,y) dx} \int_{-\infty}^{+\infty} x \cdot f_{X,Y}(x,y) dx \quad (48)$$

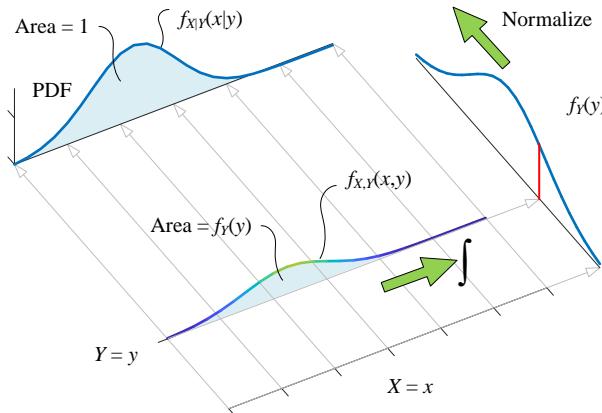


图 23. 联合概率 PDF $f_{X,Y}(x,y)$ 和条件概率 PDF $f_{X|Y}(x|y)$ 的关系， X 和 Y 均为连续随机变量

8.5 连续随机变量：条件方差

本节介绍如何求连续随机变量的条件方差。

条件方差 $\text{var}(Y|X=x)$

在给定 $X=x$ 条件下，条件方差 $\text{var}(Y|X=x)$ (conditional variance of Y given $X=x$) 定义为：

$$\begin{aligned} \text{var}(Y|X=x) &= E\left\{\left(Y - E(Y|X=x)\right)^2 | x\right\} \\ &= \int_y \left(y - E(Y|X=x)\right)^2 \cdot f_{Y|X}(y|x) dy \end{aligned} \quad (49)$$

对于连续随机变量，求条件方差也可以用 (14) 这个技巧。

条件方差 $\text{var}(X|Y=y)$

条件方差 $\text{var}(X|Y=y)$ 定义为：

$$\begin{aligned}\text{var}(X|Y=y) &= \mathbb{E}\left\{\left(X - \mathbb{E}(X|Y=y)\right)^2|y\right\} \\ &= \int_x \left(X - \mathbb{E}(X|Y=y)\right)^2 \cdot f_{X|Y}(x|y) dx\end{aligned}\quad (50)$$

有了以上理论基础，本书第 12 章将以二元高斯分布为例，继续深入讲解条件**期望**和条件**方差**。

8.6 连续随机变量：以鸢尾花为例

以鸢尾花为例：条件**期望** $E(X_2 | X_1 = x_1)$ 、条件**方差** $\text{var}(X_2 | X_1 = x_1)$

图 24 (a) 所示为条件概率 PDF $f_{X_2|X_1}(x_2 | x_1)$ 随花萼长度、花萼宽度变化曲面。本书前文提过 $f_{X_2|X_1}(x_2 | x_1)$ 也是一个二元函数。这个二元函数的重要特点有两个：

$$\begin{aligned}f_{X_2|X_1}(x_2 | x_1) &\geq 0 \\ \int_{x_2} f_{X_2|X_1}(x_2 | x_1) dx_2 &= 1\end{aligned}\quad (51)$$

正如图 24 (a) 所示，阴影区域的面积为 1。

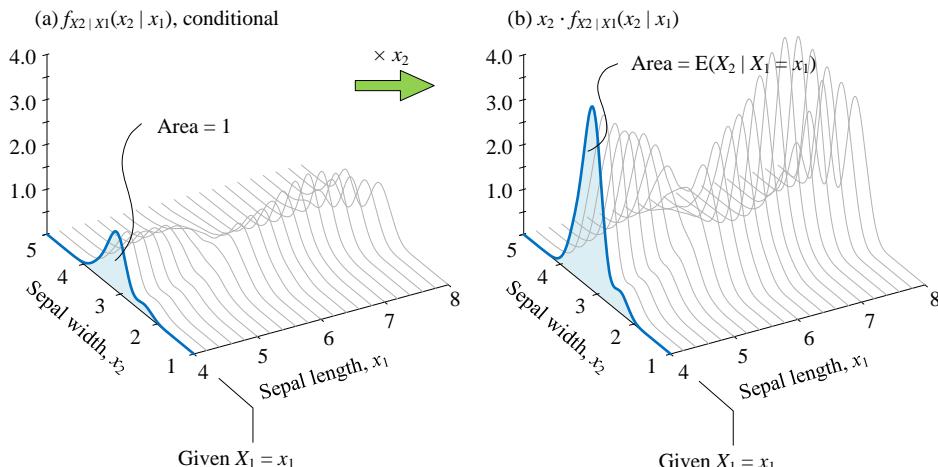


图 24. $f_{X_2|X_1}(x_2 | x_1)$ 条件概率密度三维等高线和平面等高线，不考虑分类

根据 (47)，为了计算条件**期望** $E(X_2 | X_1 = x_1)$ ，我们需要计算 $x_2 \cdot f_{X_2|X_1}(x_2 | x_1)$ 和 x_2 围成图像的面积，即图 24 (b) 阴影部分面积：

$$\mathbb{E}(X_2 | X_1 = x_1) = \int_{x_2} x_2 \cdot \underbrace{f_{X_2|X_1}(x_2 | x_1)}_{\text{Conditional}} dx_2 \quad (52)$$

然后，我们可以计算鸢尾花宽度平方的条件期望 $\mathbb{E}(X_2^2 | X_1 = x_1)$ ：

$$\mathbb{E}(X_2^2 | X_1 = x_1) = \int_{x_2} x_2^2 \cdot \underbrace{f_{X_2|X_1}(x_2 | x_1)}_{\text{Conditional}} dx_2 \quad (53)$$

然后，可以利用技巧求得条件方差 $\text{var}(X_2 | X_1 = x_1)$ ：

$$\text{var}(X_2 | X_1 = x_1) = \mathbb{E}(X_2^2 | X_1 = x_1) - \mathbb{E}(X_2 | X_1 = x_1)^2 \quad (54)$$

上式开平方得到，条件均方差 $\text{std}(X_2 | X_1 = x_1)$ 。

我们知道条件期望 $\mathbb{E}(X_2 | X_1 = x_1)$ 、条件均方差 $\text{std}(X_2 | X_1 = x_1)$ 都随着 $X_1 = x_1$ 取值变化，而且它们两个单位都是 cm。图 25 把条件期望、条件均方差整合到一幅图上。

条件期望 $\mathbb{E}(X_2 | X_1 = x_1)$ 实际上就是“回归”，给定输入条件 $X_1 = x_1$ ，求 X_2 的输出值。图 25 中黑色实线相当于“回归曲线”。

图 25 还有两条带宽 (bandwidth)，它们分别代表 $\mu_{X_2|X_1=x_1} \pm \sigma_{X_2|X_1=x_1}$ 、 $\mu_{X_2|X_1=x_1} \pm 2\sigma_{X_2|X_1=x_1}$ 。带宽随着 $X_1 = x_1$ 移动，条件均方差越大，带宽就越宽。

比较图 25、图 26，给定 $X_1 = x_1$ 条件下， X_2 上散点越集中，条件均方差 $\text{std}(X_2 | X_1 = x_1)$ 越小，比如 $X_1 = 7$ cm；相反， X_2 上散点越分散，条件均方差 $\text{std}(X_2 | X_1 = x_1)$ 越大，比如 $X_1 = 5.5$ cm。

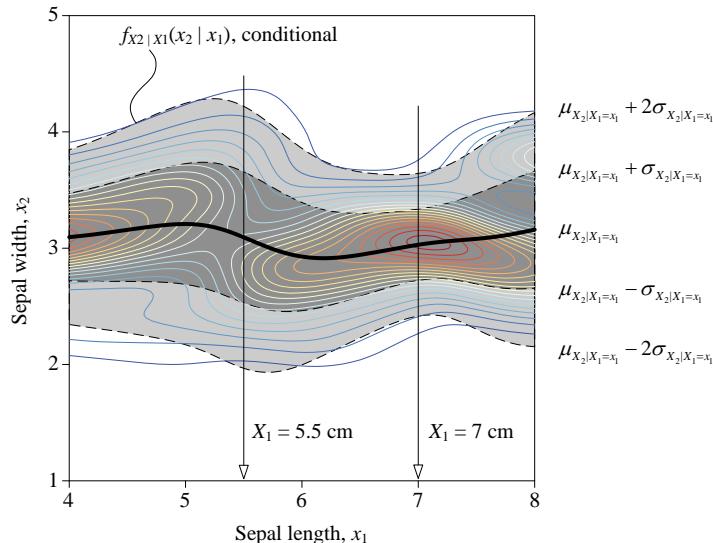


图 25. 条件期望 $\mathbb{E}(X_2 | X_1 = x_1)$ 、条件均方差 $\text{std}(X_2 | X_1 = x_1)$ 之间的关系

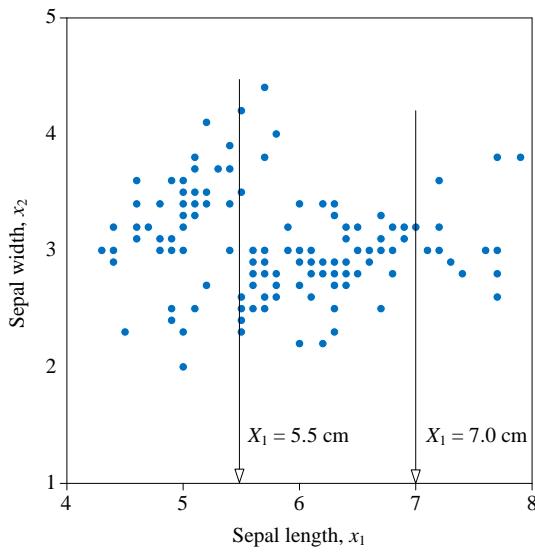


图 26. 鸢尾花数据花瓣长度、花瓣宽度散点图，不考虑分类

以鸢尾花为例：条件期望 $E(X_1 | X_2 = x_2)$ 、条件方差 $\text{var}(X_1 | X_2 = x_2)$

为了计算条件期望 $E(X_1 | X_2 = x_2)$ ，我们需要计算 $x_1 \cdot f_{X_1|X_2}(x_1 | x_2)$ 和 x_1 围成图像的面积，即图 27 (b) 阴影部分面积：

$$E(X_1 | X_2 = x_2) = \int_{-\infty}^{+\infty} x_1 \cdot \underbrace{f_{X_1|X_2}(x_1 | x_2)}_{\text{Conditional}} dx_1 \quad (55)$$

然后，我们可以计算鸢尾长度平方的条件期望 $E(X_1^2 | X_2 = x_2)$ ：

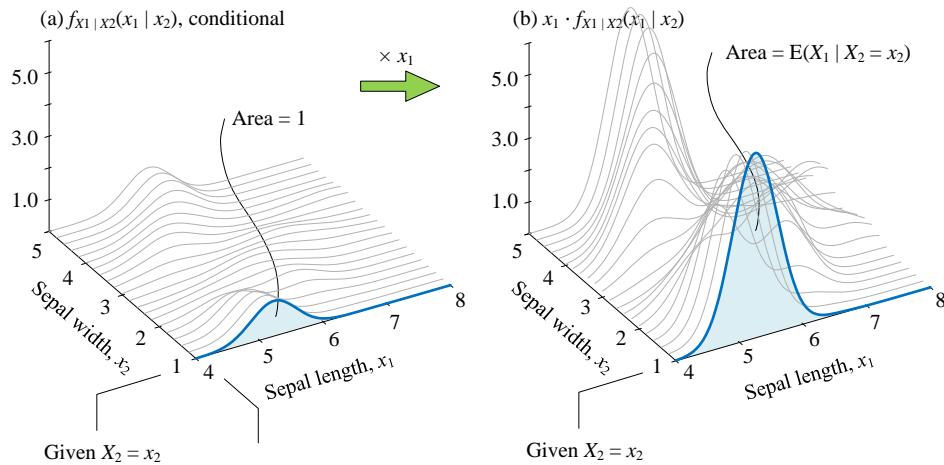
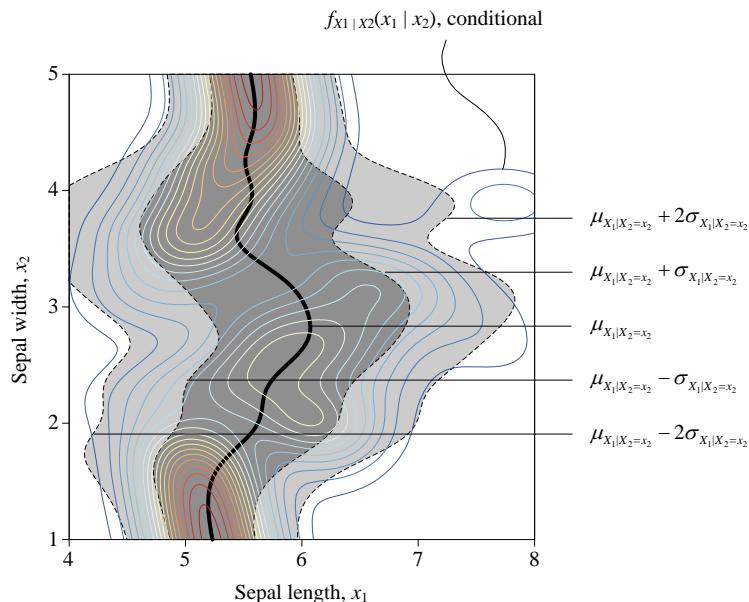
$$E(X_1^2 | X_2 = x_2) = \int_{-\infty}^{+\infty} x_1^2 \cdot \underbrace{f_{X_1|X_2}(x_1 | x_2)}_{\text{Conditional}} dx_1 \quad (56)$$

然后，可以利用技巧求得条件方差 $\text{var}(X_1 | X_2 = x_2)$ ：

$$\text{var}(X_1 | X_2 = x_2) = E(X_1^2 | X_2 = x_2) - E(X_1 | X_2 = x_2)^2 \quad (57)$$

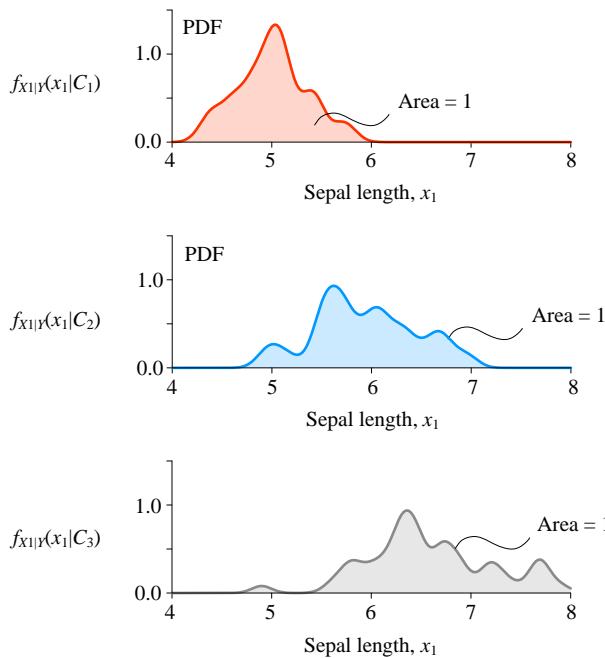
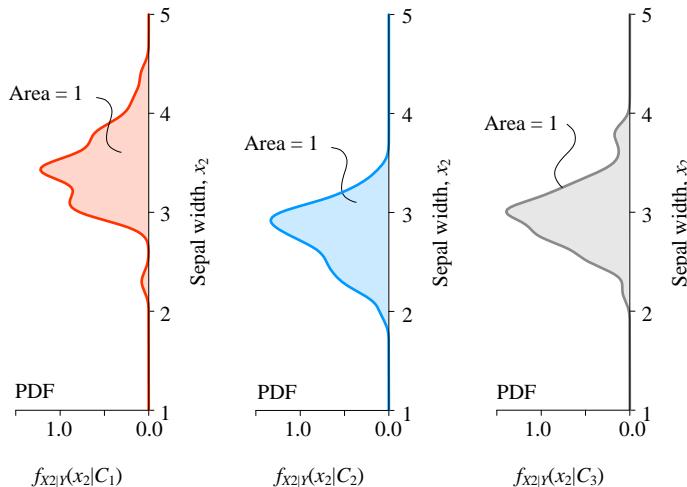
上式开平方得到条件均方差 $\text{std}(X_1 | X_2 = x_2)$ 。

我们知道条件期望 $E(X_1 | X_2 = x_2)$ 、条件标准差 $\text{std}(X_1 | X_2 = x_2)$ 都随着 $X_2 = x_2$ 取值变化，而且它们两个单位都是 cm。我们想办法把它们画在一幅图上，具体如图 28 所示。请大家自己从“回归”角度自行分析图 28。

图 27. $f_{X1|X2}(x_1 | x_2)$ 条件概率密度三维等高线和平面等高线图 28. 条件期望 $E(X_1 | X_2 = x_2)$ 、条件标准差 $\text{std}(X_1 | X_2 = x_2)$ 之间的关系

以鸢尾花为例，考虑标签

同理，我们可以计算给定标签条件下，鸢尾花萼长度（图 29）、萼片宽度（图 30）的条件期望、条件方差等。请大家自己完成这几个数值计算。

图 29. 给定鸢尾花标签 Y , 花萼长度的条件概率密度, 连续随机变量图 30. 给定鸢尾花标签 Y , 花萼宽度的条件概率密度, 连续随机变量

8.7 再谈如何分割 1

本书前文介绍过, 概率分布无非就是各种方式将样本空间概率值 1 进行“切片、切块”、“切丝、切条”。本节从这个视角总结本书这个话题讲解的主要内容。

一元

一元随机变量在一个维度上切割“1”。如果随机变量 X 离散，如图 31 (a) 所示，概率值 1 被分割成若干份，每一份还是“概率”。也就是说一元离散随机变量概率质量函数 PMF $p_X(x)$ 对应概率值。 $p_X(x)$ 对应的数学运算是 Σ 。图 31 (a) 中所有概率值之和为 1：

$$\sum_x p_X(x) = 1 \quad (58)$$

如果随机变量 X 连续，如图 31 (b) 所示， X 则对应概率密度函数 PDF $f_X(x)$ 。 $f_X(x)$ 积分结果才是概率值，因此 $f_X(x)$ 对应的数学运算符为 \int 。

$f_X(x)$ 和横轴围成的面积为 1，对应样本空间概率值“1”：

$$\int_x f_X(x) = 1 \quad (59)$$

图 31 (b) 中连续随机变量 X 的取值范围是实数轴的一个区间。图 31 (c) 中连续随机变量 X 的取值范围是整个实数轴。图 31 (c) 中， $f_X(x)$ 和整个横轴围成的面积为 1。

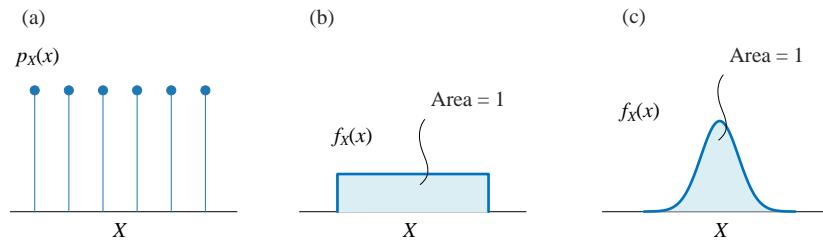


图 31. 一元随机变量

二元

二元随机变量 (X_1, X_2) 在两个维度上对样本空间进行分割。

如图 32 (a) 所示，如果 X_1 和 X_2 都是离散随机变量，概率质量函数 $p_{X_1, X_2}(x_1, x_2)$ 本身还是概率值。 $p_{X_1, X_2}(x_1, x_2)$ 二重求和的结果为 1：

$$\sum_{x_1} \sum_{x_2} p_{X_1, X_2}(x_1, x_2) = 1 \quad (60)$$

大家试图调换求和顺序时，要格外小心。并不是所有的多重求和都可以任意调换求和先后顺序。

而 $p_{X_1, X_2}(x_1, x_2)$ 偏求和便得到边缘概率质量函数 $p_{X_1}(x_1)$ 、 $p_{X_2}(x_2)$ ：

$$\begin{aligned} \sum_{x_2} p_{X_1, X_2}(x_1, x_2) &= p_{X_1}(x_1) \\ \sum_{x_1} p_{X_1, X_2}(x_1, x_2) &= p_{X_2}(x_2) \end{aligned} \quad (61)$$

如图 33 所示，二元随机变量偏求和将某个变量“消去”，这相当于折叠。

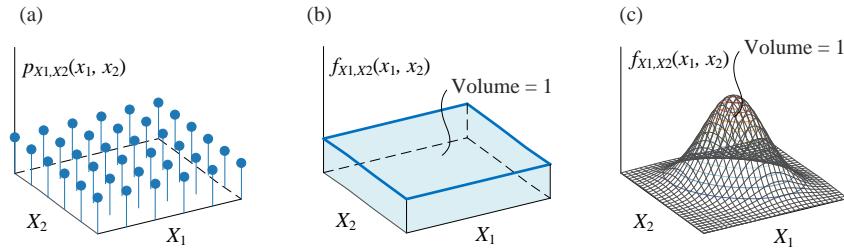


图 32. 二元随机变量

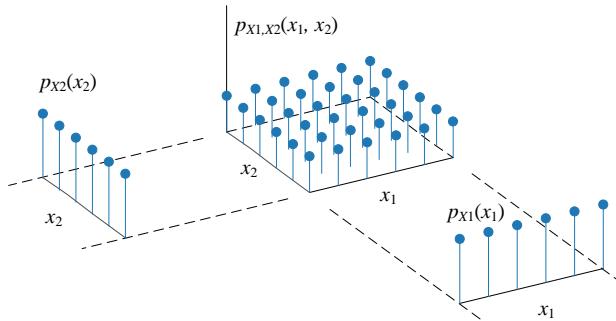


图 33. 二元随机变量偏求和，折叠某一变量

如图 32 (b) 所示，如果 X_1 和 X_2 都是连续随机变量，概率密度函数 $f_{X1,X2}(x_1, x_2)$ 如下二重积分的结果为 1：

$$\iint_{x_1 x_2} f_{X1,X2}(x_1, x_2) dx_1 dx_2 = 1 \quad (62)$$

这相当于图 32 (b) 中几何体和水平面围成的几何图形的体积为 1。如图 32 (c) 所示， X_1 和 X_2 的取值范围也可以是整个水平面，即 \mathbb{R}^2 。

$f_{X1,X2}(x_1, x_2)$ 偏积分边缘概率密度函数 $f_{X1}(x_1)$ 、 $f_{X2}(x_2)$ ：

$$\begin{aligned} \int_{x_2} f_{X1,X2}(x_1, x_2) dx_2 &= f_{X1}(x_1) \\ \int_{x_1} f_{X1,X2}(x_1, x_2) dx_1 &= p_{X2}(x_2) \end{aligned} \quad (63)$$

三元

图 34 (a) 中 (X_1, X_2, X_3) 三个随机变量都是离散随机变量，每个点 (x_1, x_2, x_3) 处都有一个概率值，这些概率值可以写成概率质量函数 $p_{X1,X2,X3}(x_1, x_2, x_3)$ 这种形式。

请大家自己写出如何根据 $p_{X_1, X_2, X_3}(x_1, x_2, x_3)$ 计算 $p_{X_1, X_2}(x_1, x_2)$ 、 $p_{X_1}(x_1)$ 。

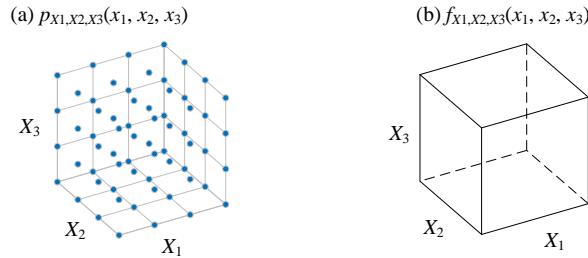


图 34. 三元随机变量

图 34 (b) 中 (X_1, X_2, X_3) 三个随机变量都是连续随机变量，整个 \mathbb{R}^3 空间中的每一点 (x_1, x_2, x_3) 处都有一个概率密度值 $f_{X_1, X_2, X_3}(x_1, x_2, x_3)$ 。这就是本书前文提到的“体密度”。也请大家自己写出如何根据 $f_{X_1, X_2, X_3}(x_1, x_2, x_3)$ 计算 $f_{X_1, X_2}(x_1, x_2)$ 、 $f_{X_1}(x_1)$ 。

图 35 所示为在 X_3 取不同值时 $X_3 = c$ ，概率密度值 $f_{X_1, X_2, X_3}(x_1, x_2, c)$ “切片”。强调一下，图 35 中 X_3 还是连续随机变量。

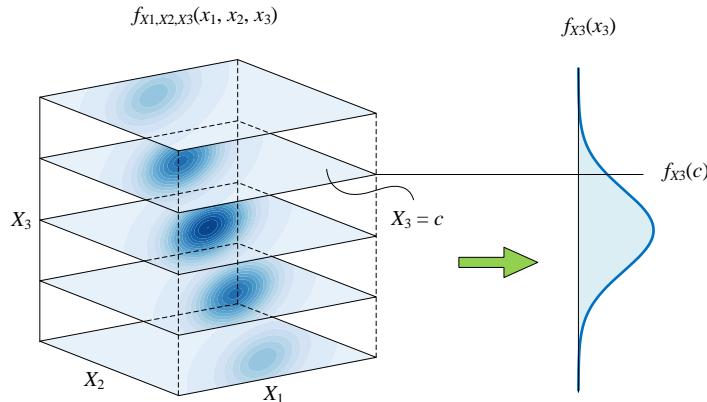


图 35. 三个随机变量都是连续随机变量

$f_{X_1, X_2, X_3}(x_1, x_2, c)$ 这个“切片”对 x_1 和 x_2 二重积分得到的是边缘概率密度 $f_{X_3}(c)$ ：

$$\int \int_{x_2, x_1} f_{X_1, X_2, X_3}(x_1, x_2, c) dx_1 dx_2 = f_{X_3}(c) \quad (64)$$

上式相当于，我们不再关心图 35 中这些切片的具体等高线，而是将其归纳为一个数值。

混合

此外，多元随机变量还可以是离散和随机变量的混合形式。一个最简单的例子就是鸢尾花数据。如图 36 所示，分类标签将鸢尾花数据分成了三层，对应 C_1 、 C_2 、 C_3 三个标签。图 36 左侧的数据构成了样本空间 Ω 。显然 C_1 、 C_2 、 C_3 互不相容，形成对样本空间 Ω 的分割。



这体现的就是本书第 3 章讲过的全概率定理。

花萼长度 X_1 、花萼宽度 X_2 都是连续随机变量，但是标签 Y 为离散随机变量。

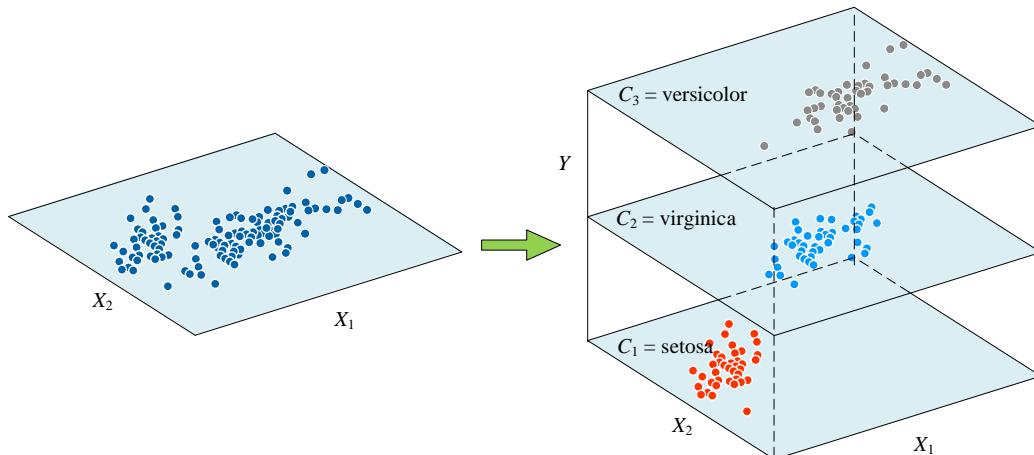


图 36. 分类标签将鸢尾花数据分层

如图 37 所示，每一类不同标签的样本数据都有其联合概率密度分布 $f_{X_1, X_2, Y}(x_1, x_2, C_1)$ 、 $f_{X_1, X_2, Y}(x_1, x_2, C_2)$ 、 $f_{X_1, X_2, Y}(x_1, x_2, C_3)$ 。

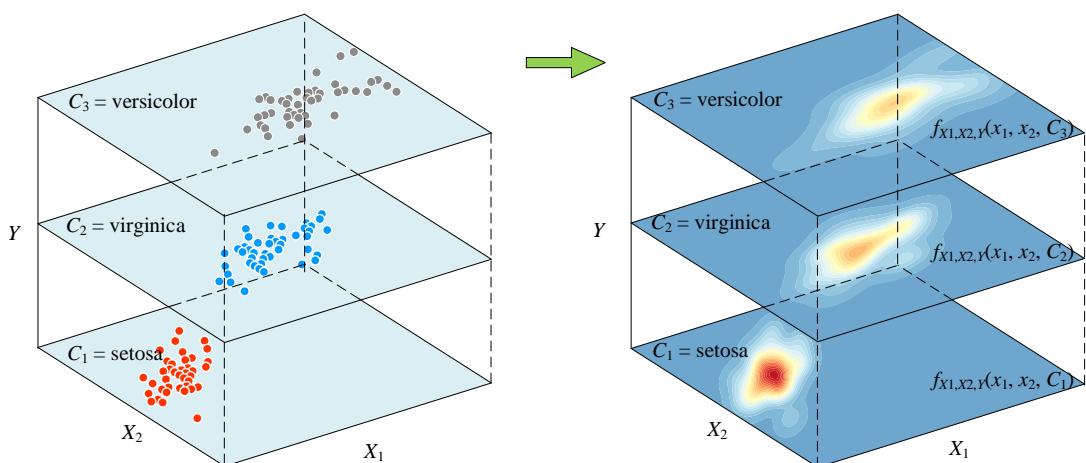


图 37. 鸢尾花数据，花萼长度 X_1 、花萼宽度 X_2 、标签 Y

图 38 所示为两个不同方向压扁 $f_{X_1, X_2, Y}(x_1, x_2, y)$ 。

$f_{X_1, X_2, Y}(x_1, x_2, C_1)$ 、 $f_{X_1, X_2, Y}(x_1, x_2, C_2)$ 、 $f_{X_1, X_2, Y}(x_1, x_2, C_3)$ 这三个平面分别二重积分得到 Y 的边缘概率：

$$\begin{aligned} \iint_{x_2 x_1} f_{X_1, X_2, Y}(x_1, x_2, C_1) dx_1 dx_2 &= p_Y(C_1) \\ \iint_{x_2 x_1} f_{X_1, X_2, Y}(x_1, x_2, C_2) dx_1 dx_2 &= p_Y(C_2) \\ \iint_{x_2 x_1} f_{X_1, X_2, Y}(x_1, x_2, C_3) dx_1 dx_2 &= p_Y(C_3) \end{aligned} \quad (65)$$

显然， $p_Y(C_1)$ 、 $p_Y(C_2)$ 、 $p_Y(C_3)$ 之和为 1。

沿着 Y 方向将 $f_{X_1, X_2, Y}(x_1, x_2, y)$ 压扁得到 $f_{X_1, X_2, Y}(x_1, x_2)$ ：

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1, X_2, Y}(x_1, x_2, C_1) + f_{X_1, X_2, Y}(x_1, x_2, C_2) + f_{X_1, X_2, Y}(x_1, x_2, C_3) \quad (66)$$

而 $f_{X_1, X_2, Y}(x_1, x_2)$ 和水平面构成的几何形体的体积为 1，即：

$$\iint_{x_2 x_1} f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 = 1 \quad (67)$$

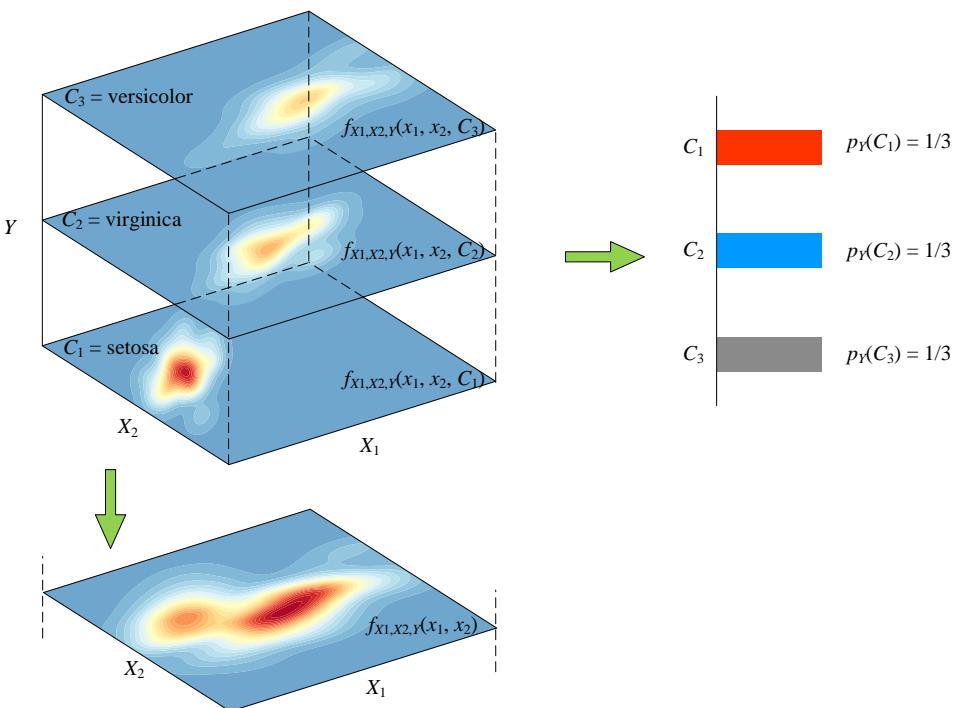


图 38. 两个不同方向压扁 $f_{X_1, X_2, Y}(x_1, x_2, y)$

此外， $f_{X_1, X_2}(x_1, x_2)$ 可以沿着不同方向进一步“压扁”得到边缘概率 $f_{X_1}(x_1)$ 、 $f_{X_2}(x_2)$ ：

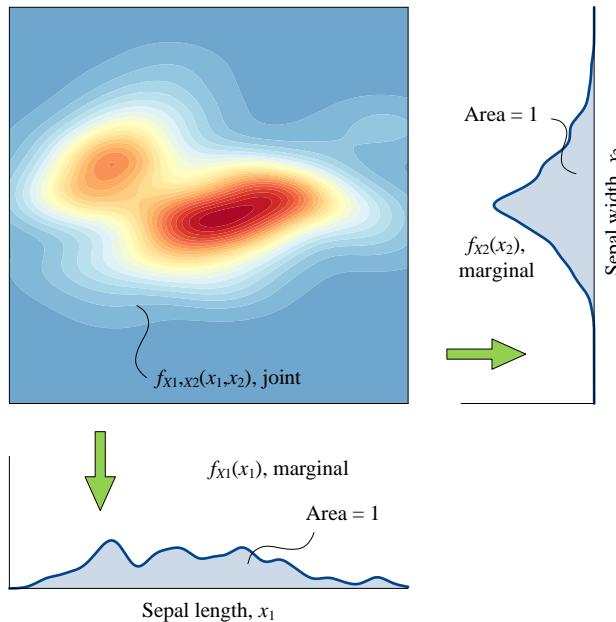
$$\int_{x_2} f_{X_1, X_2}(x_1, x_2) dx_2 = f_{X_1}(x_1) \quad (68)$$

$$\int_{x_1} f_{X_1, X_2}(x_1, x_2) dx_1 = f_{X_2}(x_2)$$

$f_{X_1}(x_1)$ 、 $f_{X_2}(x_2)$ 和 x_1 、 x_2 轴围成的面积也都是 1：

$$\int_{x_1} f_{X_1}(x_1) dx_1 = 1 \quad (69)$$

$$\int_{x_2} f_{X_2}(x_2) dx_2 = 1$$



总结来说，以上几种情况无非就是对 1 的“切片、切块”、“切丝、切条”。

此时，希望大家闭上眼睛想 $f_{X_1, X_2, Y}(x_1, x_2, C_1)$ 、 $f_{X_1, X_2}(x_1, x_2)$ 的时候看到的是等高线；想 $f_{X_1}(x_1)$ 看到曲线，想 $p_Y(C_1)$ 的时候看到一个数值 (1/3)。

不同的混合形式

图 39 所示为二元随机变量的不同离散、连续混合形式。图 39 (a) 两个随机变量都是连续。图 39 (b) 中 X_1 为离散随机变量， X_2 为连续随机变量；图 39 (c) 反之。图 39 (d) 中，两个随机变量都是离散随机变量。图 40 所示为三元随机变量的不同离散、连续混合形式，请大家自己分析其中子图。这实际上回答了本书第 4 章提出的问题。

在本书贝叶斯统计推断(第 20~22 章)中，大家会发现我们不再区分 PDF、PMF，概率分布函数全部统一为 $f()$ 。

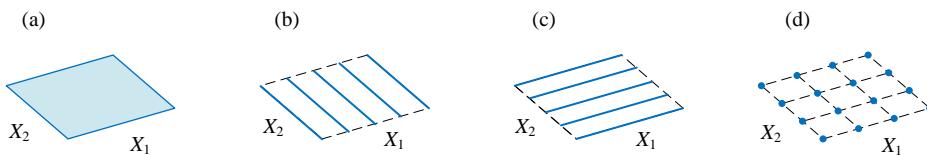


图 39. 二元随机变量，混合

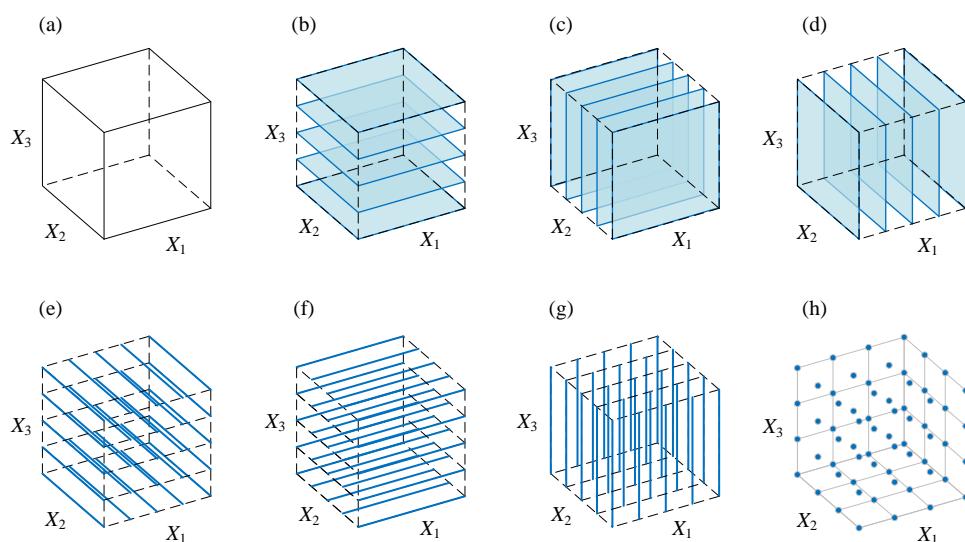


图 40. 三元随机变量，混合

条件概率：重新定义 1

条件概率其实很好理解，条件概率的“条件”就是划定“新的样本空间”，对应概率值也是 1。也就是说，把从原始样本空间中切出来的“一片、一块、一丝、一条”作为新的样本空间。

如图 41 所示，给定标签为 $Y = C_2$ 条件下，利用贝叶斯定理，条件概率可以通过下式求得：

$$f_{X_1, X_2|Y}(x_1, x_2 | C_2) = \frac{f_{X_1, X_2, Y}(x_1, x_2, C_2)}{p_Y(C_2)} \quad (70)$$

分母中的 $p_Y(C_2)$ 起到归一化的作用。 $Y = C_2$ 就是原始样本空间中切出来的“一片”。

也就是说， $f_{X_1, X_2|Y}(x_1, x_2 | C_2)$ 二重积分的结果为 1：

$$\iint_{x_2, x_1} f_{X_1, X_2|Y}(x_1, x_2 | C_2) dx_1 dx_2 = 1 \quad (71)$$

上式中这个“1”对应条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | C_2)$ 的条件—— $Y = C_2$ 。 $Y = C_2$ 就是这个条件概率的“新样本空间”。



本书第 6 章还介绍过，以鸢尾花萼长度或宽度为条件的条件概率，请大家回顾。



鸢尾花书《可视之美》介绍如何绘制本节分层等高线。

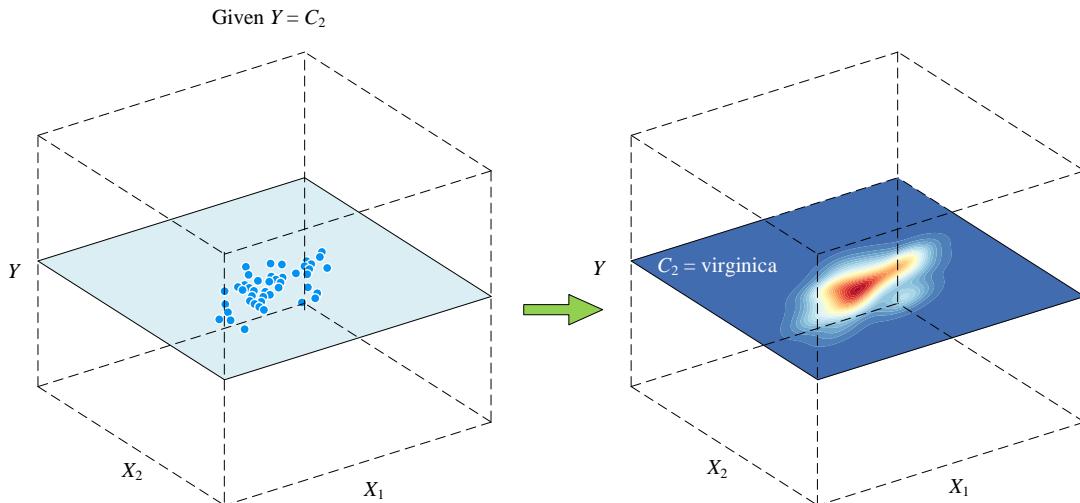


图 41. 条件概率，给定标签为 $Y = C_2$



条件期望是指在已知一些条件下，一个随机变量的期望值。同理，条件方差是指在给定某些条件下，随机变量的方差。它俩表示给定某些信息或事件之后，对随机变量的期望、方差的预测或估计。其实生活中条件期望、方差无处不在，大家多多留意。条件期望、方差在概率论、统计学和经济学等领域有广泛的应用，例如在回归分析、决策树、贝叶斯推断等中。

至此，本书“概率”板块介绍。下一版块将用五章深入介绍高斯分布，一元、二元、多元、条件高斯分布，以及协方差矩阵。

9 Univariate Gaussian Distribution 一元高斯分布

可能是应用最广泛的概率分布



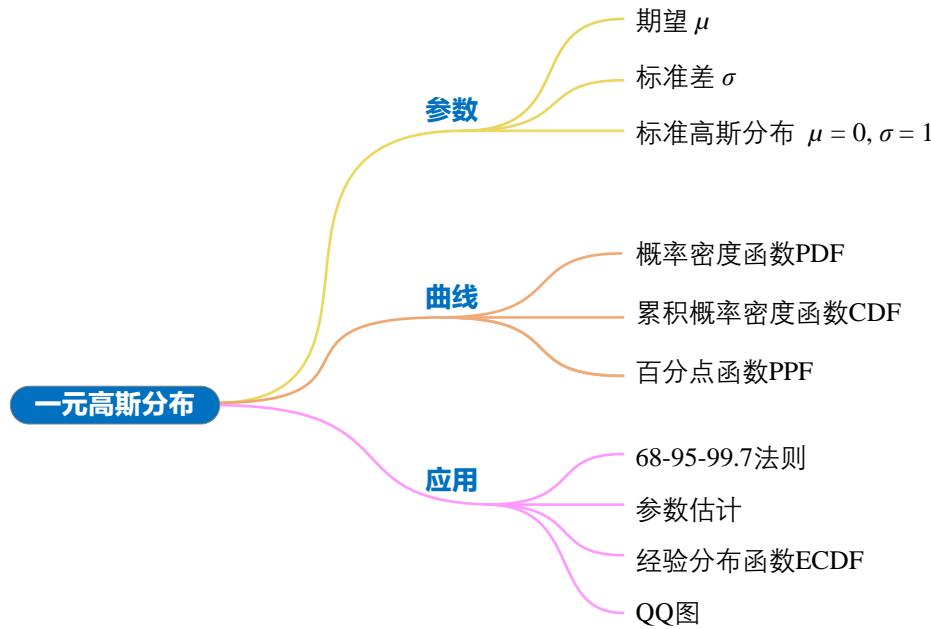
数学家站在彼此的肩膀上。

Mathematicians stand on each other's shoulders.

—— 卡尔·弗里德里希·高斯 (Carl Friedrich Gauss) | 德国数学家、物理学家、天文学家 | 1777 ~ 1855



- ◀ `matplotlib.pyplot.axhline()` 绘制水平线
- ◀ `matplotlib.pyplot.axvline()` 绘制竖直线
- ◀ `matplotlib.pyplot.contour()` 绘制等高线图
- ◀ `matplotlib.pyplot.contourf()` 绘制填充等高线图
- ◀ `numpy.ceil()` 计算向上取整
- ◀ `numpy.copy()` 深拷贝数组，对新生成的对象修改删除操作不会影响到原对象
- ◀ `numpy.cumsum()` 计算累积和
- ◀ `numpy.floor()` 向下取整
- ◀ `numpy.meshgrid()` 生成网格数据
- ◀ `numpy.random.normal()` 生成满足高斯分布的随机数
- ◀ `scipy.stats.norm.cdf()` 高斯分布累积分布函数 CDF
- ◀ `scipy.stats.norm.pdf()` 高斯分布概率密度函数 PDF
- ◀ `scipy.stats.norm.ppf()` 高斯分布百分点函数 PPF



9.1 一元高斯分布：期望值决定位置，标准差决定形状

回顾上一章介绍**一元高斯分布** (univariate normal distribution)，其概率密度函数 PDF 如下：

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (1)$$

其中， μ 为期望值， σ 为标准差。

期望值

一元高斯分布概率密度函数的形状为中间高两边低的钟形，其 PDF 最大值位于 $x = \mu$ 。

本书前文提过，一元高斯分布的概率密度函数以 $x = \mu$ 为轴左右对称，曲线向左右两侧远离 $x = \mu$ 呈逐渐均匀下降趋势，曲线两端与横轴 $y = 0$ 无限接近，但永不相交。

图 1 所示为 μ 对一元高斯分布 PDF 曲线位置的影响。

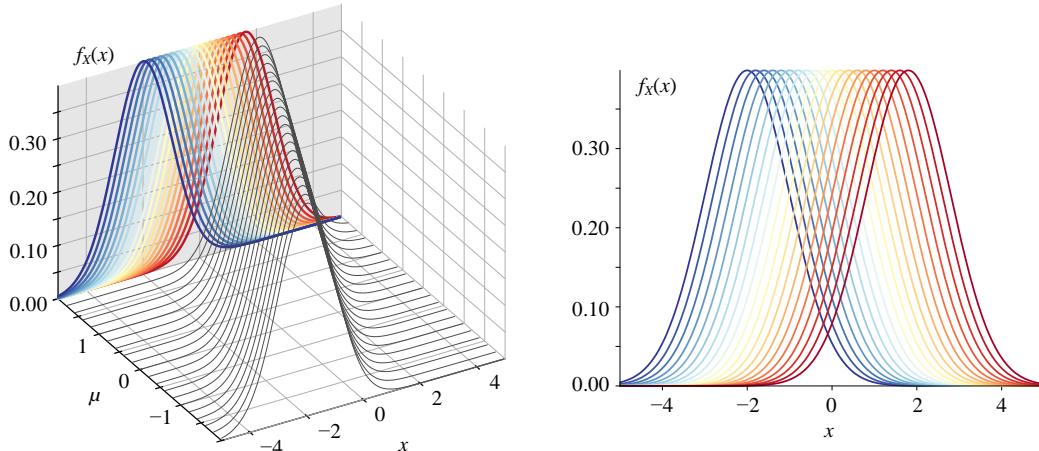


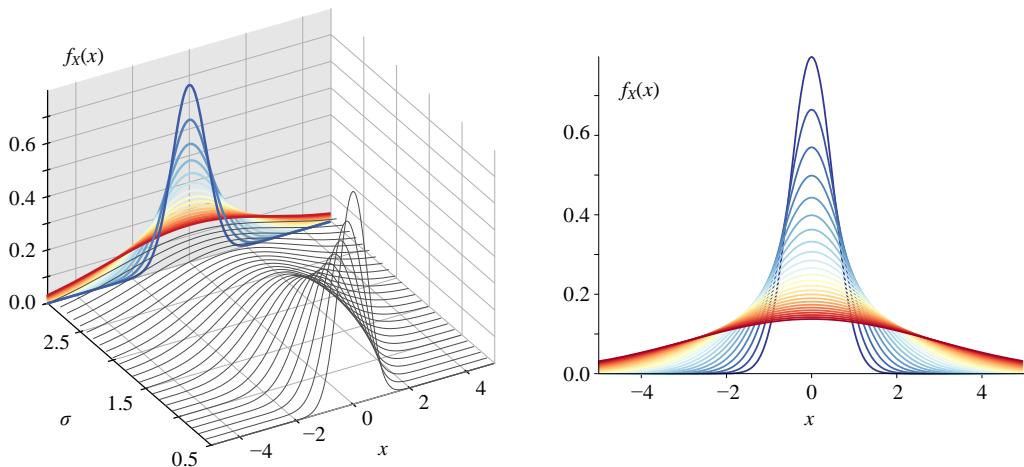
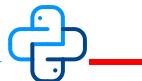
图 1. μ 对一元高斯分布 PDF 曲线位置的影响

标准差

σ 也称为高斯分布的形状参数， σ 越大，曲线越扁平；反之， σ 越小，曲线越瘦高。

从数据角度来讲， σ 描述数据分布的离散程度。 σ 越大，数据分布越分散， σ 越小，数据分布越集中。图 2 所示为 σ 对一元高斯分布 PDF 曲线形状影响。

本书前文强调过，期望值、标准差的单位和随机变量的单位相同。因此，直方图、概率密度图上常常出现 $\mu \pm \sigma$ 、 $\mu \pm 2\sigma$ 、 $\mu \pm 3\sigma$ 等等。

图 2. σ 对一元高斯分布 PDF 曲线形状影响

Bk5_Ch09_01.py 绘制图 1。请大家修改代码自行绘制图 2。代码自定义函数计算一元高斯分布概率密度，大家也可以使用 `scipy.stats.norm.pdf()` 函数获得一元高斯分布密度函数值。



在 Bk5_Ch09_01.py 基础上，我们用 Streamlit 制作了一个应用，大家可以改变 μ 、 σ 参数值，观察一元高斯 PDF 曲线变化。请大家参考 Streamlit_Bk5_Ch09_01.py。

9.2 累积概率密度：对应概率值

一元高斯分布的累积概率密度函数 CDF：

$$F_x(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt \quad (2)$$

上式也可以用误差函数 `erf()` 表达：

$$F_x(x) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right] \quad (3)$$



《数学要素》第 18 章介绍过误差函数，请大家回顾。

期望值

图 3 所示为 μ 对一元高斯分布 CDF 曲线位置的影响。随着 x 不断靠近 $-\infty$, CDF 取值不断接近于 0, 但不等于 0; 反之, 随着 x 不断靠近 $+\infty$, CDF 取值不断接近于 1, 但不等于 1。

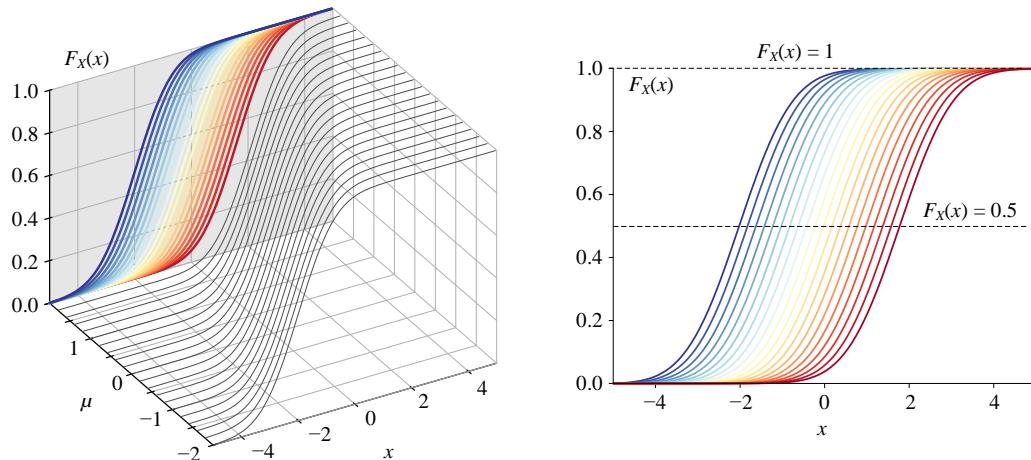


图 3. μ 对一元高斯分布 CDF 曲线位置的影响

标准差

图 4 所示为 σ 对一元高斯分布 CDF 曲线形状影响。 σ 越小, CDF 曲线越陡峭; σ 越大, 越平缓。从另外一个角度看一元高斯分布 CDF 曲线, 它将位于实数轴 $(-\infty, +\infty)$ 之间的 x 转化为 $(0, 1)$ 之间的某个值, 而这个值恰好对应一个概率。

注意, 图 1、图 2 的纵轴对应概率密度值, 而图 3、图 4 的纵轴对应概率值。也就是说, 一元概率密度函数积分的结果为概率值。

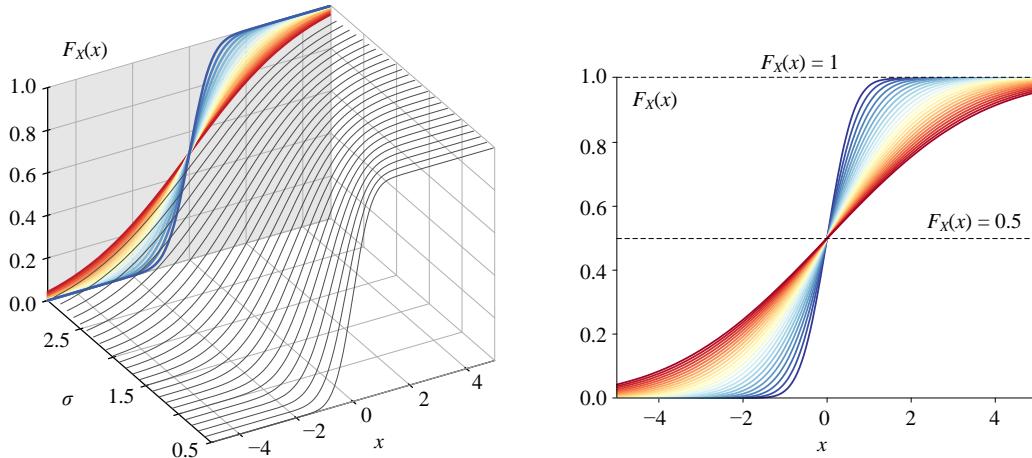


图 4. σ 对一元高斯分布 CDF 曲线形状影响



Bk5_Ch08_02.py 绘制图 3 和图 4。

PDF vs CDF

图 5 比较标准正态分布 $N(0, 1)$ 的 PDF 和 CDF 曲线。虽然两条曲线画在同一幅图上，它们的 y 轴数值的含义完全不同。对于 PDF 曲线，它的 y 轴数值代表概率密度，并不是概率值。而 CDF 曲线的 y 轴数值则代表概率值。

给定一点 x ，图 5 中背景为浅蓝色区域面积对应 $F_x(x) = \int_{-\infty}^x f_x(t) dt$ ，也就是 CDF 曲线的高度值。下一节还会继续讲解标准正态分布。

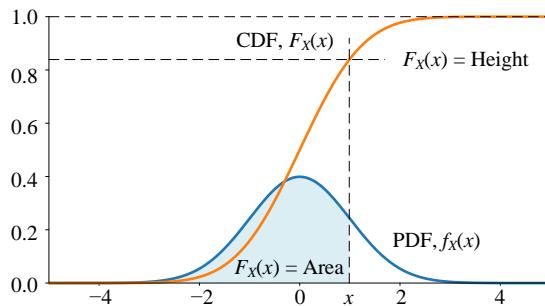


图 5. 比较标准正态分布的 PDF 和 CDF 曲线

百分点函数 PPF

我们把 Percent-Point Function (PPF) 直译为“**百分点函数**”。实际上，百分点函数 PPF 是 **CDF 逆函数** (inverse CDF)。

如图 6 所示，给定 x ，我们可以通过 CDF 曲线得到累积概率值 $F_x(x) = p$ 。而 PPF 曲线则正好相反，给定概率值 p ，通过 PPF 曲线得到 x ，即 $F_x^{-1}(p) = x$ 。在 SciPy 中，正态分布的 CDF 函数为 `scipy.stats.norm.cdf()`，对应的 PPF 函数为 `scipy.stats.norm.ppf()`。

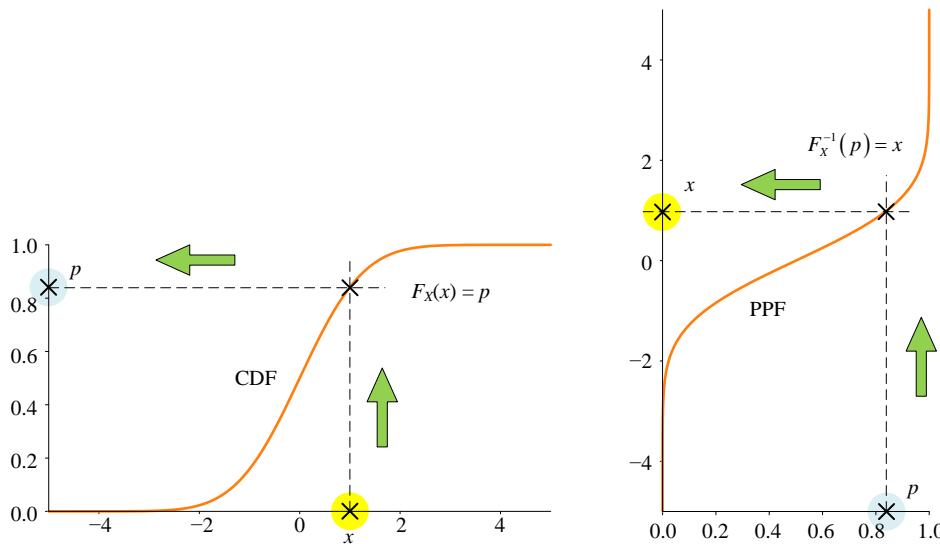


图 6. CDF 曲线和 PPF 曲线之间关系

9.3 标准高斯分布：期望为 0，标准差为 1

当 $\mu = 0$ 且 $\sigma = 1$ 时，高斯分布为**标准正态分布** (standard normal distribution)，记做 $N(0, 1)$ 。

本节用 Z 表示服从标准正态分布的连续随机变量，而 Z 的实数取值用 z 代表。因此，标准正态分布的 PDF 函数为：

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (4)$$

可以写成， $Z \sim N(0, 1)$ 。

图 7 (a) 所示为标准正态分布 PDF 曲线。特别地， $Z = 0$ 时，标准高斯分布的概率密度值为：

$$f_Z(0) = \frac{1}{\sqrt{2\pi}} \approx 0.39894 \quad (5)$$

这个值经常近似为 0.4。再次强调，0.4 这个值虽然也代表可能性，但是它不是概率值，是概率密度值。

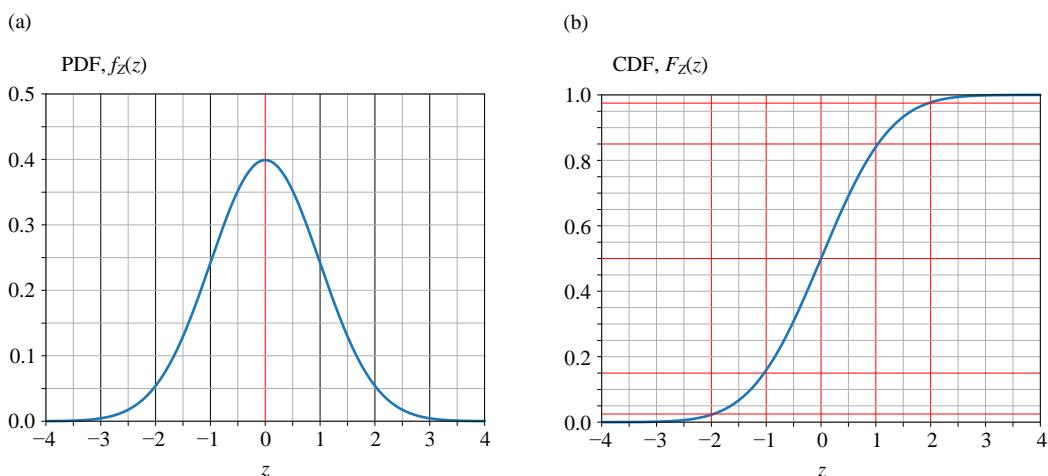


图 7. 标准高斯分布 PDF 和 CDF 曲线

容易发现，当 PDF 曲线 $f_Z(z)$ 随着 z 增大而增大时 (对称轴左半边)，PDF 的增幅先是逐渐变大，曲线逐渐变陡；然后，PDF 的增幅放缓，曲线坡度逐渐变得平缓，在 $z = 0$ 曲线坡度为 0。

从一阶导数的角度来看，对于 PDF 曲线对称轴左半边，一阶导数值大于 0，直到 $z = 0$ 处，即均值 μ 处，一阶导数值为 0。

然而，这段曲线 z 从负无穷增大到 0 时，二阶导数先为正，中间穿过 0，然后为负值。

PDF 曲线二阶导数为 0 正好对应 $\mu \pm \sigma$ 这两点，这两点正是 PDF 曲线的拐点。

图 8 所示为标准正态分布 $N(0, 1)$ 的 CDF、PDF、PDF 一阶导数、PDF 二阶导数这四条曲线。其中，黑色 \times 对应 PDF 曲线的最大值处。红色 \times 对应 PDF 曲线拐点。请大家仔细分析这四幅图像中曲线的变化趋势。

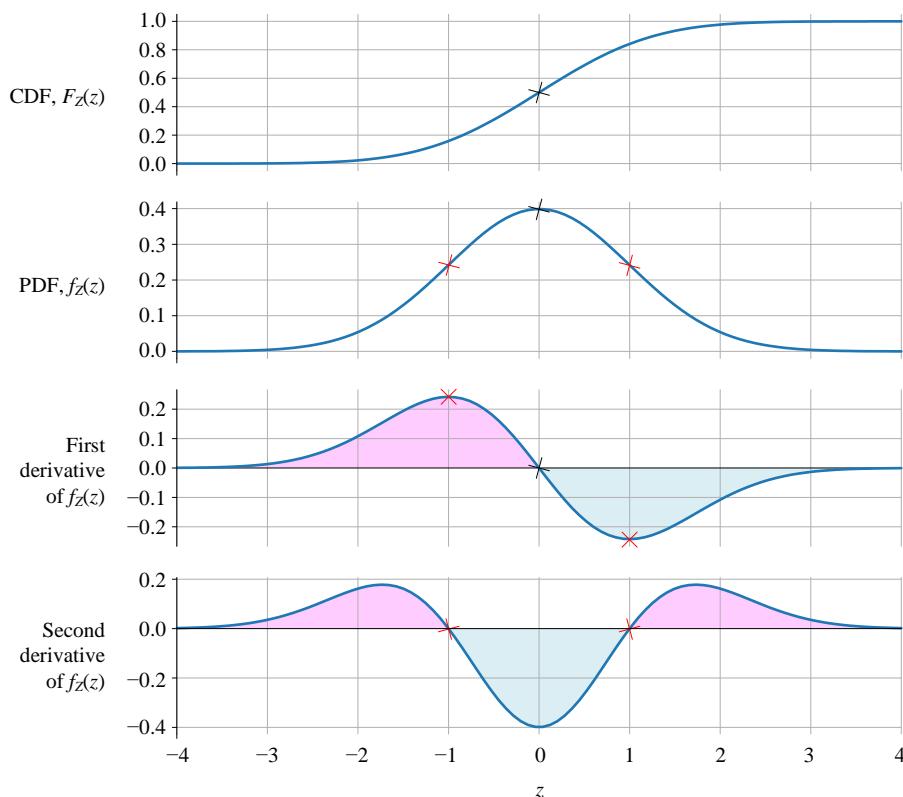


图 8. 四条曲线：标准正态分布 CDF、PDF、PDF 一阶导数、PDF 二阶导数

Z 分数：一种以标准差为单位的度量尺度

Z 分数 (Z-score)，也叫**标准分数** (standard score) 是样本值 x 与平均数 μ 的差再除以标准差 σ 的结果，对应的运算为：

$$z = \frac{x - \mu}{\sigma} \quad (6)$$

上述过程也叫做数据的**标准化** (standardize)。样本数据的 Z 分数构成的分布有两个特点：a) 平均等于 0；b) 标准差等于 1。

从距离的角度来看，(6) 代表数据点 x 和均值 μ 之间的距离为 z 倍标准差 σ 。

注意，本书前文强调，标准差和 x 具有相同的单位，而 (6) 分式消去了单位，这说明 Z 分数**无单位** (unitless)。

⚠ 注意，本书把“normalize”翻译为“归一化”，它表示将一组数据转化为 $[0, 1]$ 区间的数值。线性代数中，**向量单位化** (vector normalization) 指的是将非零向量转化成 L^2 模为 1 的单位向量。很多资料混用“standardize”和“normalize”，请大家注意区分。

图 9 所示为标准正态分布随机变量 z 值和 PDF $f_Z(z)$ 的对应关系。图 10 所示为标准正态分布 z 值到 CDF 值的映射关系。图 11 所示为 PPF 值到标准正态分布 z 值的映射关系。本章前文介绍过，CDF 和 PPF 互为反函数。

图 12 所示为标准正态分布中，不同 z 值对应的四类面积。我们一般会在 **Z 检验** (Z test) 中用到这个表。

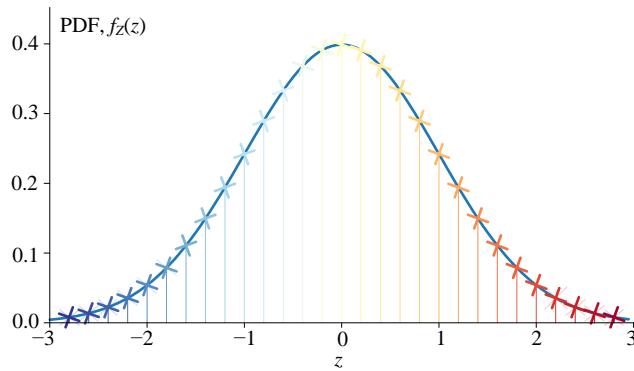


图 9. 标准正态分布 z 和 PDF 的对应关系

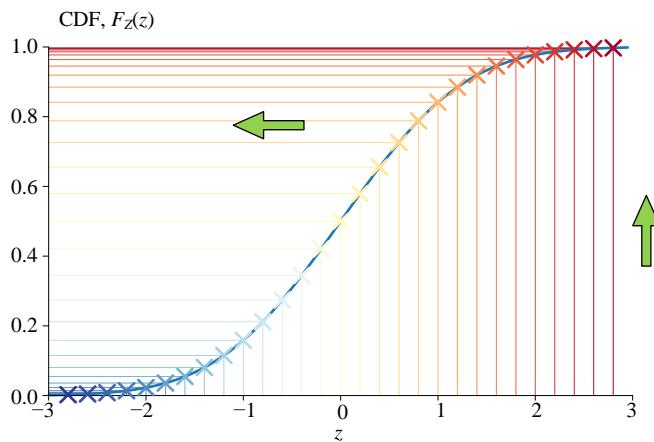
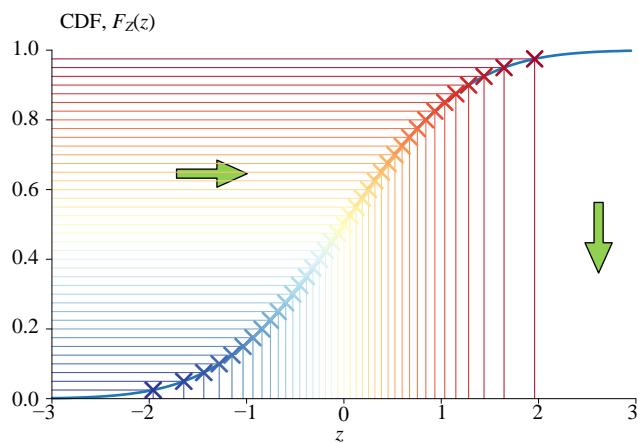
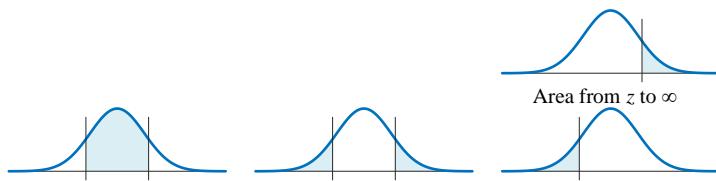
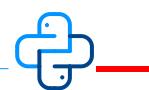


图 10. 标准正态分布 z 和 CDF 值的映射关系

图 11. 标准正态分布 z 和 PPF 值的映射关系

z	Area from $-z$ to z	Area from $-\infty$ to $-z$ and z to ∞	Area from $-\infty$ to $-z$
1.64485	0.9	0.1	0.05
1.66959	0.905	0.095	0.0475
1.69540	0.91	0.09	0.045
1.72238	0.915	0.085	0.0425
1.75069	0.92	0.08	0.04
1.78046	0.925	0.075	0.0375
1.81191	0.93	0.07	0.035
1.84526	0.935	0.065	0.0325
1.88079	0.94	0.06	0.03
1.91888	0.945	0.055	0.0275
1.95996	0.95	0.05	0.025
2.00465	0.955	0.045	0.0225
2.05375	0.96	0.04	0.02
2.10836	0.965	0.035	0.0175
2.17009	0.97	0.03	0.015
2.24140	0.975	0.025	0.0125
2.32635	0.98	0.02	0.01
2.43238	0.985	0.015	0.0075
2.57583	0.99	0.01	0.005
2.80703	0.995	0.005	0.0025
3.29053	0.999	0.001	0.0005

图 12. 标准正态分布中，不同 z 值对应的四类面积

Bk5_Ch08_03.py 绘制本节之前大部分图像。

以鸢尾花数据为例

前文提过，Z 分数可以看成一种标准化的“距离度量”。原始数据的 Z 分数代表距离均值若干倍的标准差偏移。比如，某个数据点的 Z 分数为 3，说明这个数据距离均值 3 倍标准差偏移。Z 分数的正负表达偏移的方向。比如，如果某个样本点的 Z 分数为 -2，这意味着该样本点位于均值左侧，距离均值 2 倍标准差。

有了 Z 分数，不同分布、不同单位的样本数据有了可比性。图 13 所示为鸢尾花样本数据四个特征的 Z 分数。

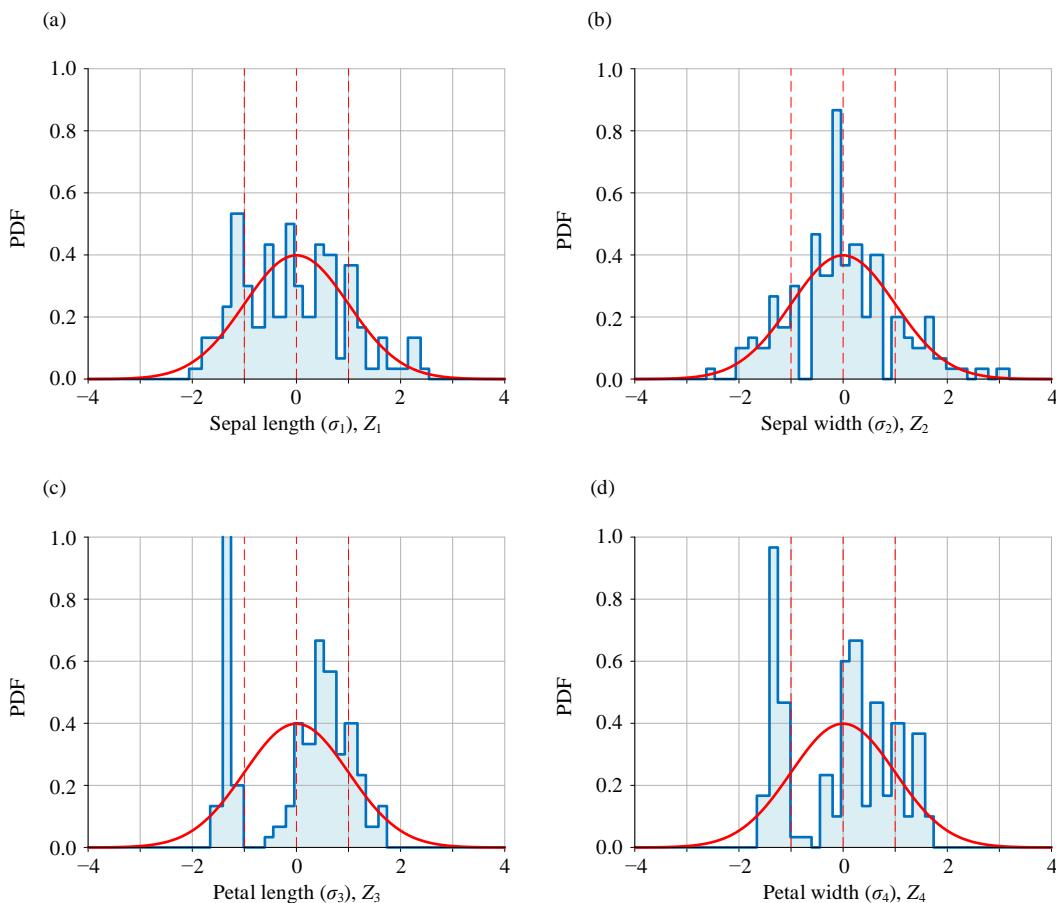


图 13. 鸢尾花四个特征的 z 分数，标准差距离

9.4 68-95-99.7 法则

一元高斯分布有所谓的 **68-95-99.7 法则** (68-95-99.7 Rule)，具体是指一组近乎满足正态分布的样本数据，约 68.3%、95.4% 和 99.7% 样本位于距平均值正负 1 个、2 个和 3 个标准差范围之内。

标准正态分布 $N(0, 1)$

以标准正态分布 $N(0, 1)$ 为例，整条标准正态分布曲线和横轴包裹的面积为 1。

如图 14 (a) 所示， $[-1, 1]$ 区间内，标准正态分布和横轴包裹的区域面积约为 0.68，即 68%。

如图 14 (b) 所示， $[-2, 2]$ 区间对应的阴影区域面积约为 0.95，即 95%。

如图 14 (c) 所示， $[-3, 3]$ 区间对应的阴影区域面积约为 0.997，即 99.7%。

写成具体的概率运算：

$$\begin{aligned} \Pr(-1 \leq Z \leq 1) &\approx 0.68 \\ \Pr(-2 \leq Z \leq 2) &\approx 0.95 \\ \Pr(-3 \leq Z \leq 3) &\approx 0.997 \end{aligned} \tag{7}$$

图 15 所示为标准正态分布 CDF 曲线上 68-95-99.7 法则对应的高度。

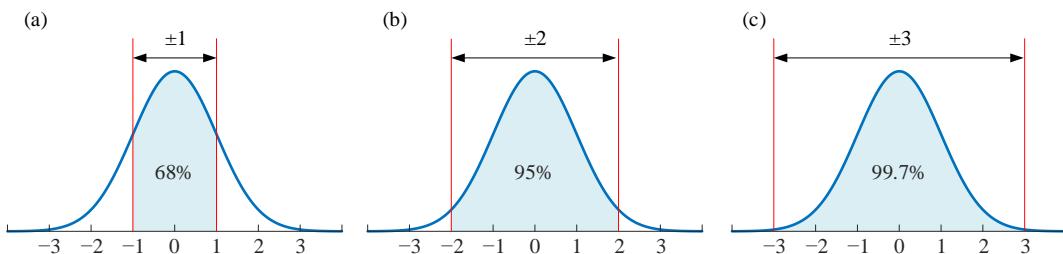


图 14. 68-95-99.7 法则，标准正态分布 PDF

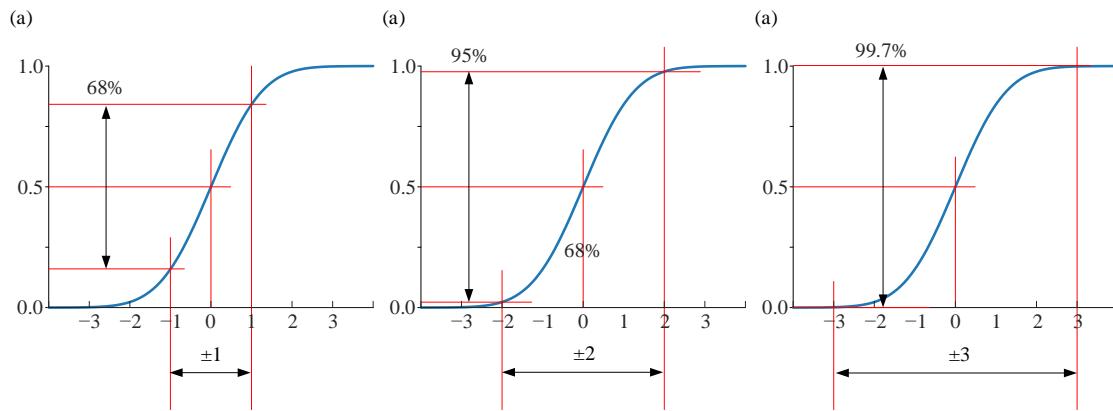


图 15. 68-95-99.7 法则，标准正态分布 CDF

正态分布 $N(\mu, \sigma^2)$

图 16 所示为一般正态分布 $N(\mu, \sigma^2)$ 中 68-95-99.7 法则对应的位置：

$$\begin{aligned} \Pr(\mu - \sigma \leq X \leq \mu + \sigma) &\approx 0.68 \\ \Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) &\approx 0.95 \\ \Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) &\approx 0.997 \end{aligned} \quad (8)$$

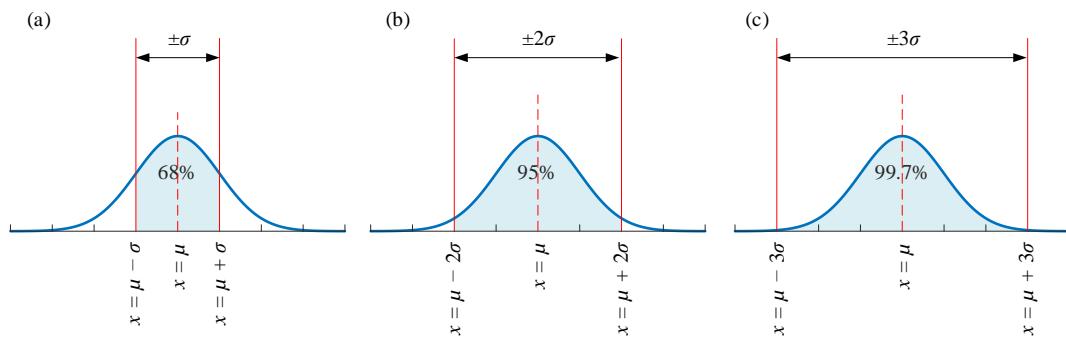


图 16. 68-95-99.7 法则，一般正态分布

和分位数的关系

图 17 所示为 68-95-99.7 法则和四分位、十分位、二十分位、百分位关系。

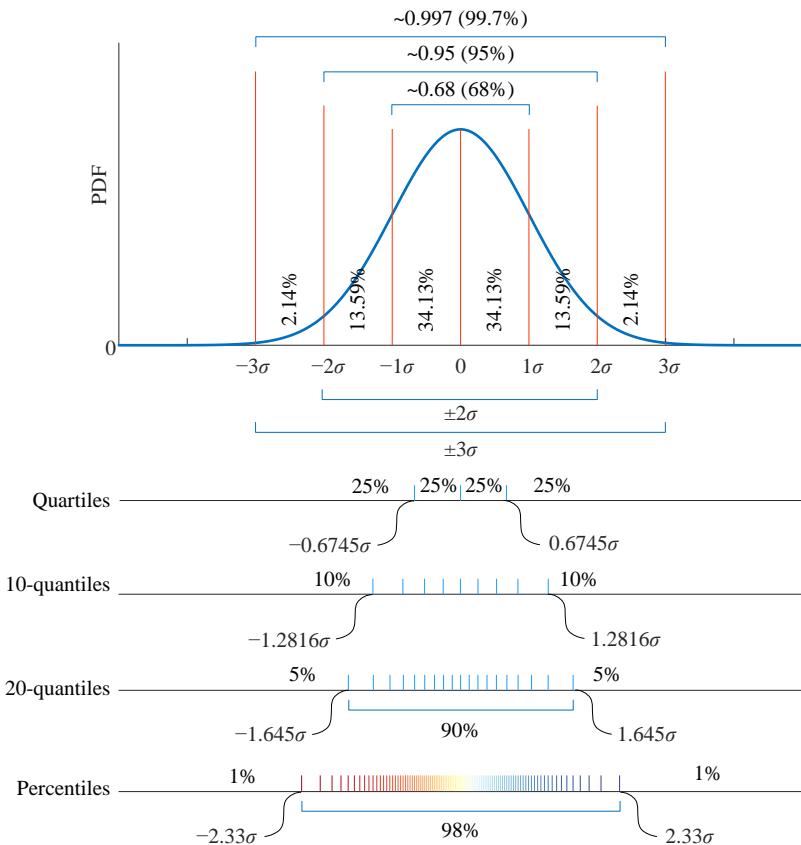


图 17. 68-95-99.7 法则和四分位、十分位、二十分位、百分位关系。注意图中并不区分总体标准差 σ 和样本标准差 s

随机数

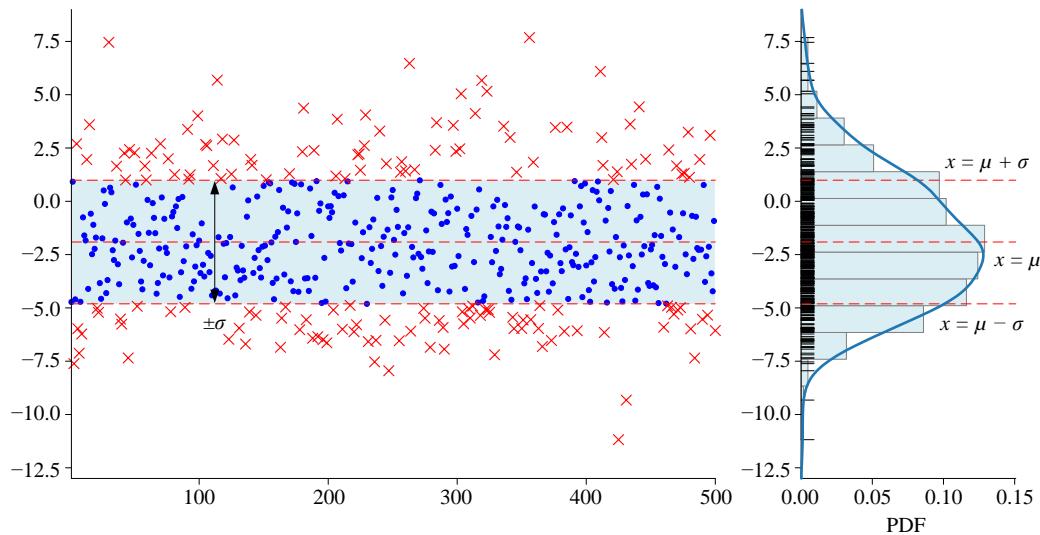
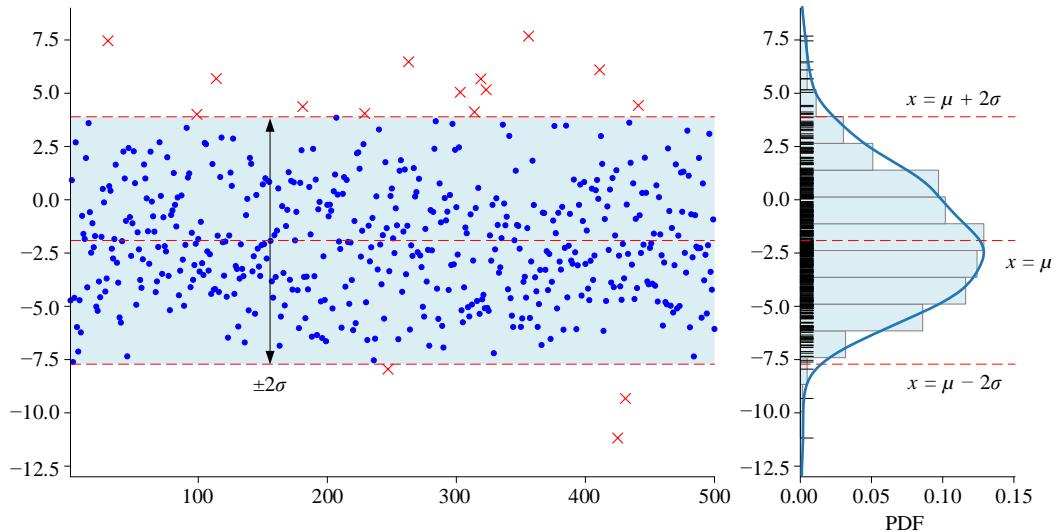
如果随机数服从一元高斯分布 $N(\mu, \sigma^2)$, 在 $[\mu - \sigma, \mu + \sigma]$ 这个 $\mu \pm \sigma$ 区间内, 应该约有 68% 的随机数。如图 18 所示, 样本一共有 500 个随机数, 约 340 个 ($= 500 \times 68\%$) 在 $\mu \pm \sigma$ 之内, 约 160 个在 $\mu \pm \sigma$ 之外。

在 $[\mu - 2\sigma, \mu + 2\sigma]$ 这个 $\mu \pm 2\sigma$ 区间内, 应该约有 95% 的随机数。如图 19 所示, 样本数还是 500 个, 约 475 个 ($= 500 \times 95\%$) 在 $\mu \pm 2\sigma$ 之内, 约 25 个在 $\mu \pm 2\sigma$ 之外。

68-95-99.7 法则可以帮助大家直观地理解一元高斯分布的形态和特征, 即大部分数据集中在均值周围, 而远离均值的数据较为稀少。如果一组数据中存在明显偏离均值多个标准差的数据点, 就有可能是异常值或者离群值, 需要进一步检查和分析。



鸢尾花书《数据有道》将专门介绍如何发现离群值。

图 18. 500 个随机数和 $\mu \pm \sigma$ 图 19. 500 个随机数和 $\mu \pm 2\sigma$ 

Bk5_Ch08_04.py 绘制图 18 和图 19。

9.5 用一元高斯分布估计概率密度

概率密度估计：参数估计

在数据科学和机器学习中，**概率密度估计** (probability density estimation) 是经常遇到的一个问题。简单来说，概率密度估计就是从离散的样本数据中估计得到连续的概率密度函数曲线。白话讲，找到一条 PDF 曲线尽可能贴合样本数据分布。

一元高斯分布 PDF 只需要两个参数——均值 (μ)、标准差 (σ)。有些时候，一元高斯分布是估计某个特定特征样本数据分布的一个不错且很便捷的选择。

以鸢尾花数据为例

举个例子，样本数据中花萼长度的均值为 $\mu_1 = 5.843$ ，标准差为 $\sigma_1 = 0.825$ 。

注意， μ_1 和 σ_1 的单位均为厘米。

有了这两个参数，我们可以用一元高斯分布估计鸢尾花花萼长度随机变量 X_1 概率密度函数：

$$f_{X_1}(x) = \frac{1}{\sqrt{2\pi} \times 0.825} \exp\left(-\frac{1}{2}\left(\frac{x-5.843}{0.825}\right)^2\right) \quad (9)$$

类似地，我们可以用一元高斯分布估计鸢尾花其他三个特征的 PDF。这样便得到图 20 所示四条 PDF 曲线。

有了概率密度函数，我们可以回答这样的问题，比如鸢尾花的花萼长度在 [4, 6] cm 区间的概率大概多少？利用定积分运算就可以得到量化结果。

给定样本数据，采用一元高斯分布估计单一特征概率密度函数很简单；但是，这种估算方法对应的问题也很明显。

比如，图 20 (a) 和 (b) 告诉我们用高斯分布描述鸢尾花花萼长度和花萼宽度样本数据分布似乎还可以接受。

但是，比较图 20 (c) 和 (d) 中直方图和高斯分布，显然高斯分布不适合描述鸢尾花花瓣长度和宽度样本数据分布。



本书第 18 章将利用核密度估计解决这一问题。

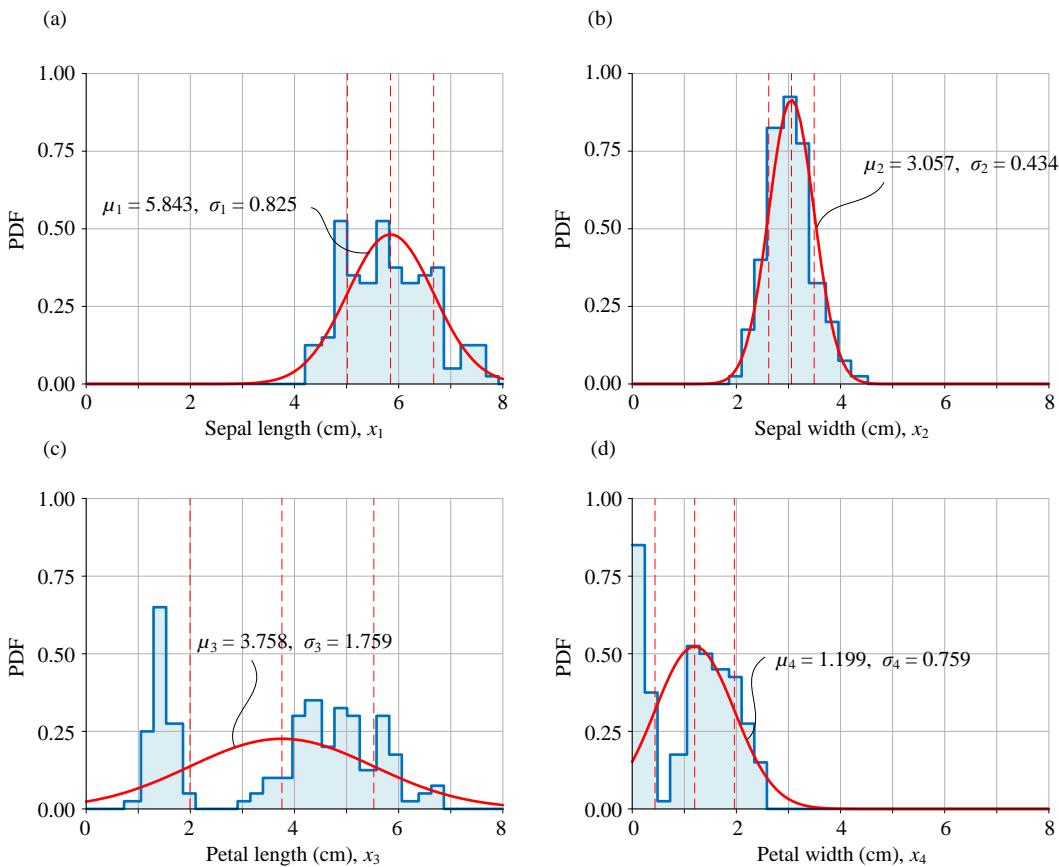


图 20. 比较概率密度直方图和高斯一元分布 PDF

9.6 经验分布函数

经验分布函数 (empirical cumulative distribution function, ECDF) 是用来描述一组样本数据分布情况的统计工具。ECDF 将样本数据按照大小排序，并计算每个数据点对应的累计比例，形成一个类似阶梯函数的曲线，横坐标表示数据的取值，它的纵坐标则表示小于等于横坐标的数据比例。

具体来说，如果有 n 个样本，ECDF 是在所有 n 个数据点上都跳跃 $1/n$ 的阶跃函数。

显然，累积概率函数是一个双射函数。从函数角度来讲，**双射** (bijection) 指的是每一个输入值都有正好一个输出值，并且每一个输出值都有正好一个输入值。

ECDF 常常用来与理论 CDF 分布函数进行比较，以检验样本数据是否符合某种假设的分布。

图 21 比较鸢尾花不同特征样本数据的 ECDF 和对应的高斯分布 CDF 曲线。

逆经验累积分布函数 (inverse empirical cumulative distribution function, inverse ECDF) 是 ECDF 的逆函数。图 22 比较逆经验累积分布函数和高斯分布 PPF 曲线。

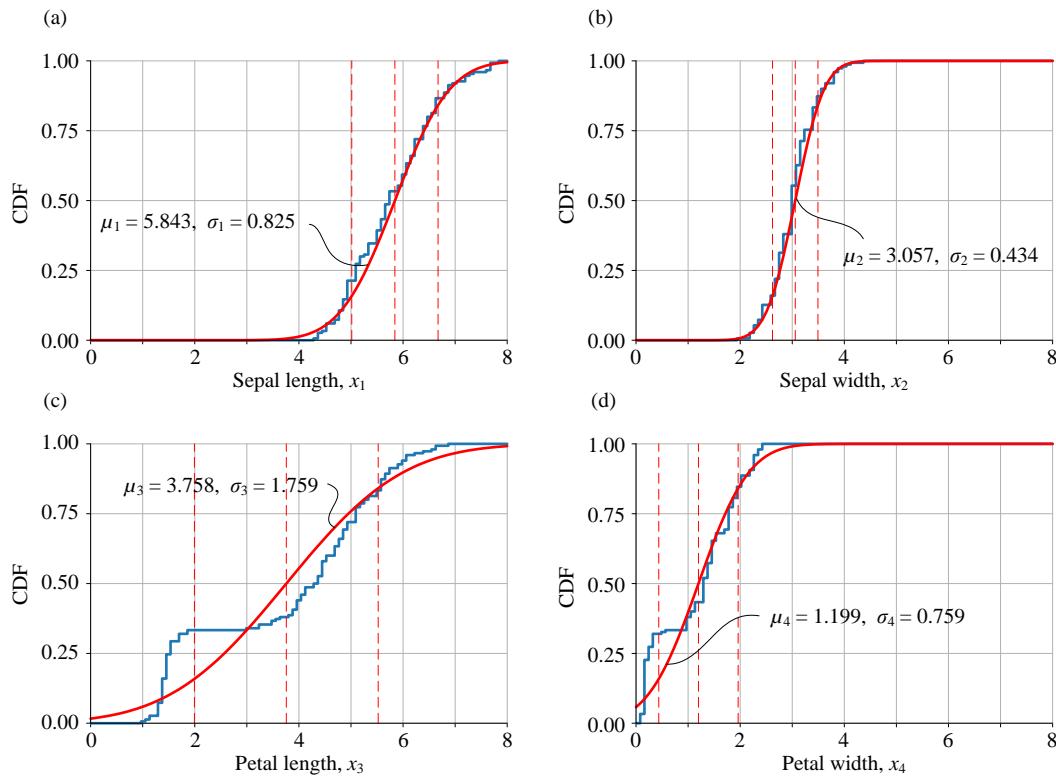


图 21. 比较 ECDF 和高斯 CDF

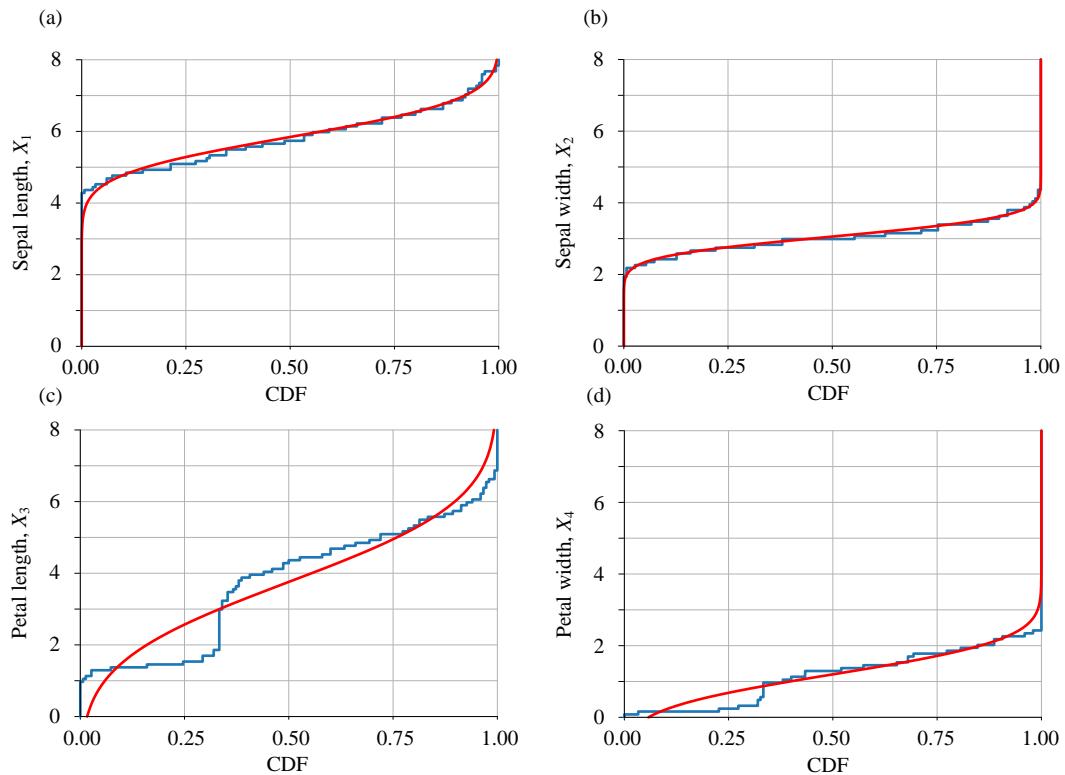


图 22. 逆经验累积分布函数和高斯 PPF

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

9.7 QQ 图：分位-分位图

QQ 图 (quantile-quantile plot, QQ plot) 中的 Q 代表分位数，常用于检查数据是否符合某个分布的统计图形。QQ 图是散点图，横坐标一般为假定分布 (比如标准正态分布) 分位数，纵坐标为待检验样本的分位数。

图 23 所示为 QQ 图原理。我们首先计算每个样本 $y^{(i)}$ 对应的 ECDF 值，然后再利用标准正态分布 PPF 将 ECDF 值转化为 $x^{(i)}$ 。这样我们便获得一系列散点 $(x^{(i)}, y^{(i)})$ 。

在 QQ 图中，将假定分布和待检验样本的分位数相互对应，从而比较它们之间的相似度。如果样本符合假定分布，则 QQ 图呈现出一条近似于直线的对角线，如果不符，则呈现出偏离直线的曲线形状。

QQ 图的横坐标一般是正态分布，当然也可以是其他分布。

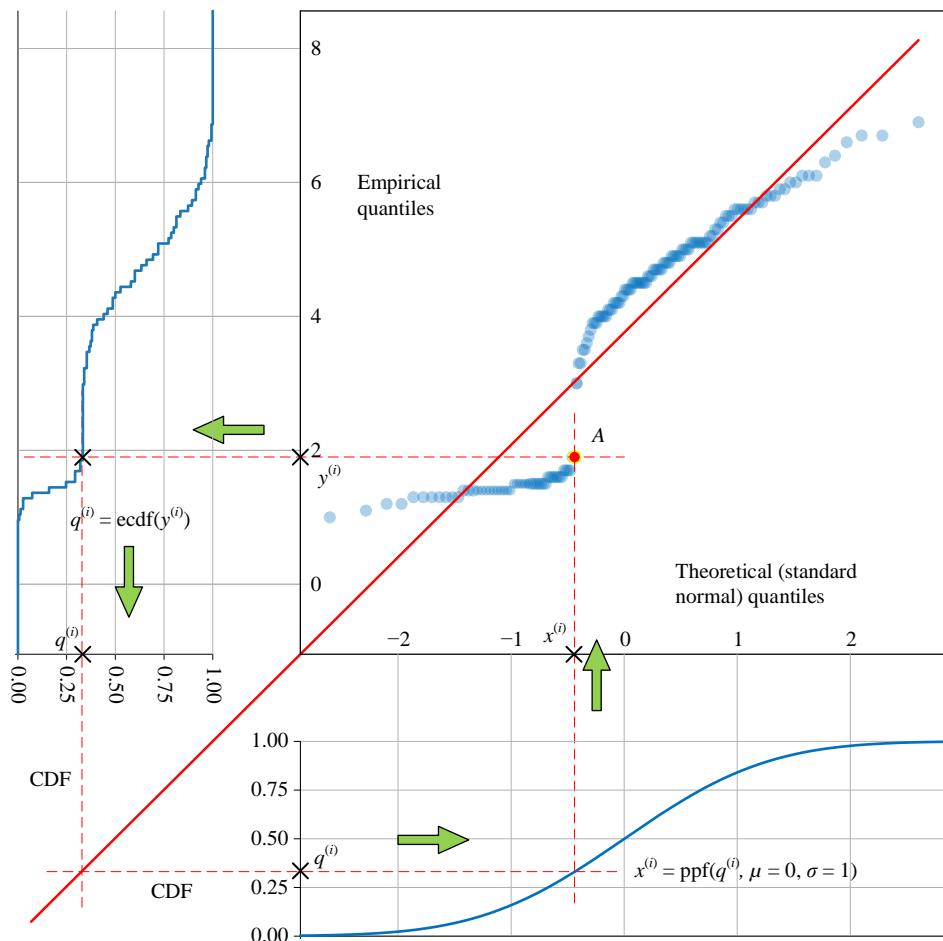


图 23. QQ 图原理，横轴为正态分布

以鸢尾花数据为例

图 24 所示为鸢尾花数据四个特征样本数据的 QQ 图。通过观察这四幅图像，大家应该能够看出那个特征的数据分布更类似（贴合）正态分布。这和图 20、图 21 得出的结论相同。换个角度来看，QQ 图实际上就是图 21 的另外一种可视化方案。

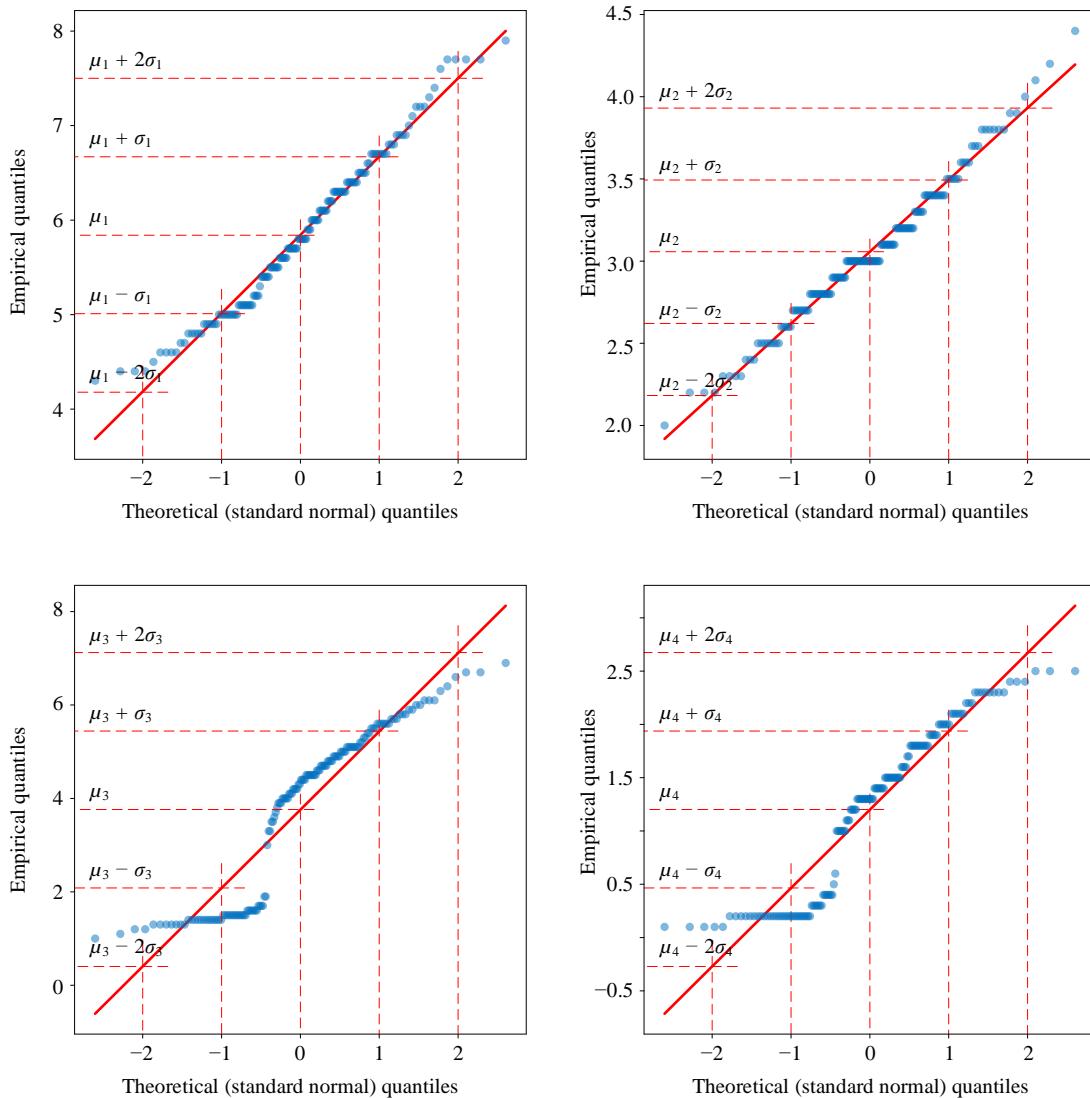


图 24. 鸢尾花数据四个特征样本数据的 QQ 图



Bk5_Ch08_05.py 绘制 8.5 ~ 8.7 节大部分图像。

特殊分布的 QQ 图特征

图 25 所示为几种常见特殊分布对比正态分布的 QQ 图。如图 25 (a) 所示，当样本数据分布近似服从正态分布时，QQ 图中散点几乎在一条直线上。通过散点图的形态，我们还可以判断分布是否有双峰（图 25 (b)）、瘦尾（图 25 (c)）、肥尾（图 25 (d)）、左偏（图 25 (e)）、右偏（图 25 (f)）。

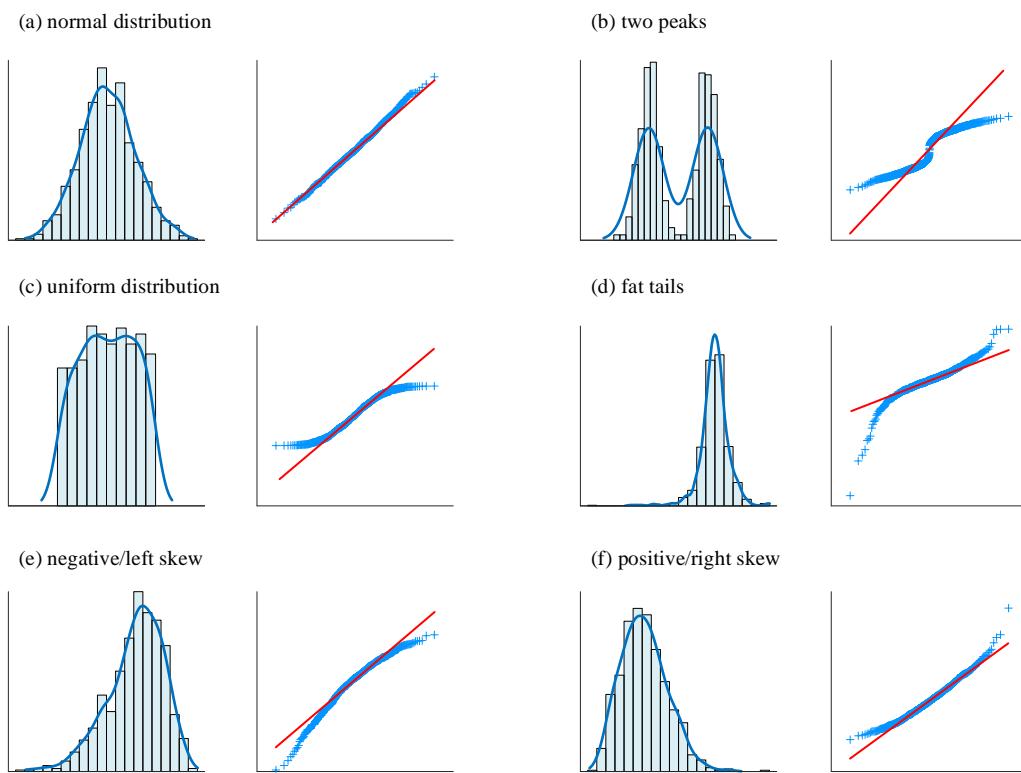


图 25. 几种特殊分布的 QQ 图特点，对比纵坐标正态分布

当然 QQ 图的横轴也可以是其他分布的 CDF。图 26 所示为横轴为均匀分布的 QQ 图，即横轴为理论均匀分布，纵轴为近似均匀分布的样本数据。

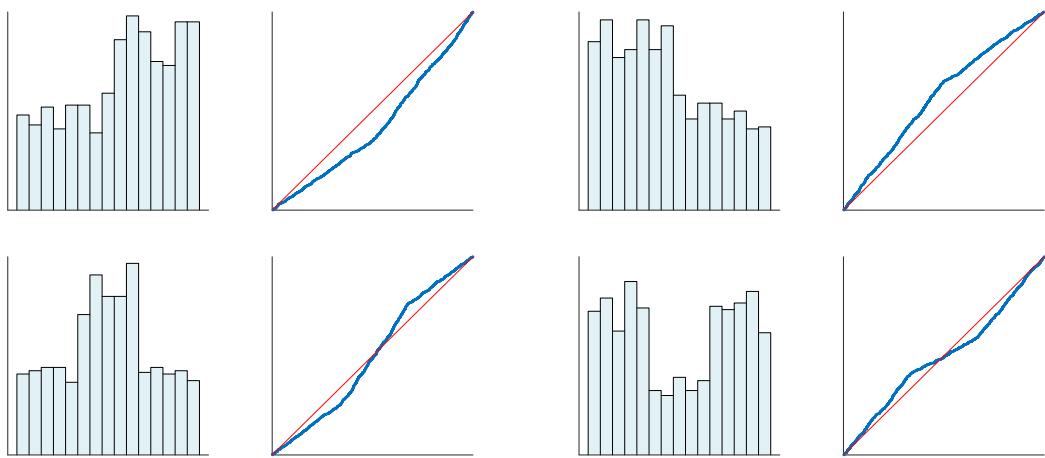


图 26. 几种特殊分布的 QQ 图特点，横轴为理论均匀分布，纵轴为近似均匀分布的样本数据

9.8 从距离到一元高斯分布

现在回过头来再看一元高斯分布的 PDF 解析式：

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (10)$$

而标准正态分布的 PDF 解析式为：

$$f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (11)$$

几何变换：平移 + 缩放

比较 (1) 和 (4)，我们容易发现满足 $N(\mu, \sigma^2)$ 的 X 可以通过“平移 (translate) + 缩放 (scale)”变成满足 $N(0, 1)$ 的 Z 。 $X \rightarrow Z$ 对应的运算为：

$$Z = \frac{X - \mu}{\sigma} \quad \begin{matrix} \text{Translate} \\ \text{---} \\ \text{Scale} \end{matrix} \quad (12)$$

相反， $Z \rightarrow X$ 对应“缩放 + 平移”：

$$X = \underbrace{Z\sigma + \mu}_{\text{Scale}} \quad \begin{matrix} \text{Translate} \\ \text{---} \\ \text{Scale} \end{matrix} \quad (13)$$

图 27 所示为满足 $N(10, 4)$ 的一元高斯分布通过“平移 + 缩放”变成标准高斯分布的过程。

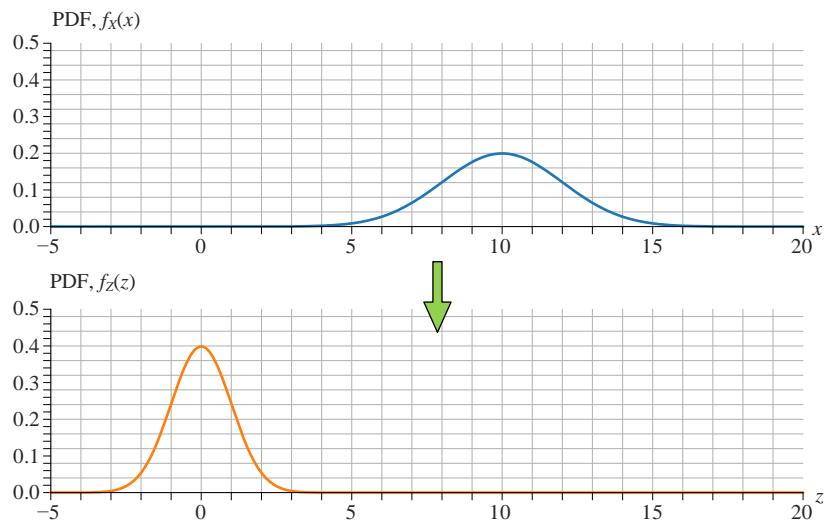


图 27. 随机变量 X 线性变换得到 Z 的过程

如图 28 所示，平移仅改变随机数的均值位置，不影响随机数的分布情况。如图 29 所示，缩放改变随机数的分布离散程度。

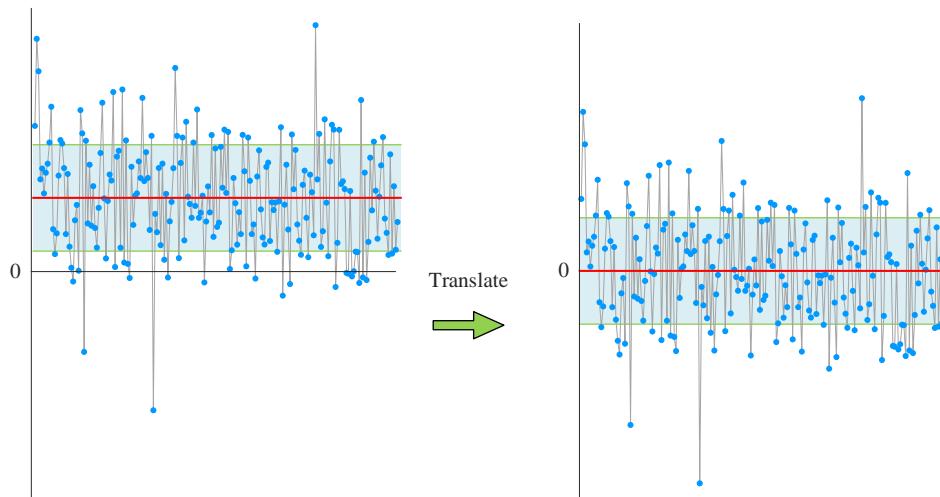


图 28. 平移

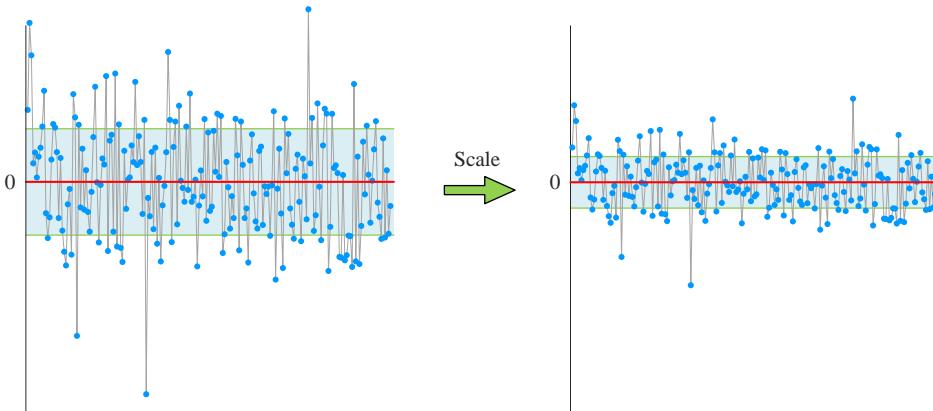


图 29. 缩放

假设 X 是连续随机变量，它的概率密度函数 PDF 为 $f_X(x)$ ，经过如下线性变换得到 Y ：

$$Y = aX + b \quad (14)$$

Y 的 PDF 为：

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \quad (15)$$

这样就解释了 (10) 和 (11) 的关系。

注意，(14) 相当于线性代数中的仿射变换。

此外，服从正态分布的随机变量，在进行线性变换后，正态性保持不变。比如， X 为服从 $N(\mu, \sigma^2)$ 的随机变量； $Y = aX + b$ 仍然服从正态分布。 Y 的均值、方差分别为：

$$\mathbb{E}(Y) = a\mu + b, \quad \text{var}(Y) = a^2\sigma^2 \quad (16)$$

图 30 所示为随机变量线性变换的示意图。

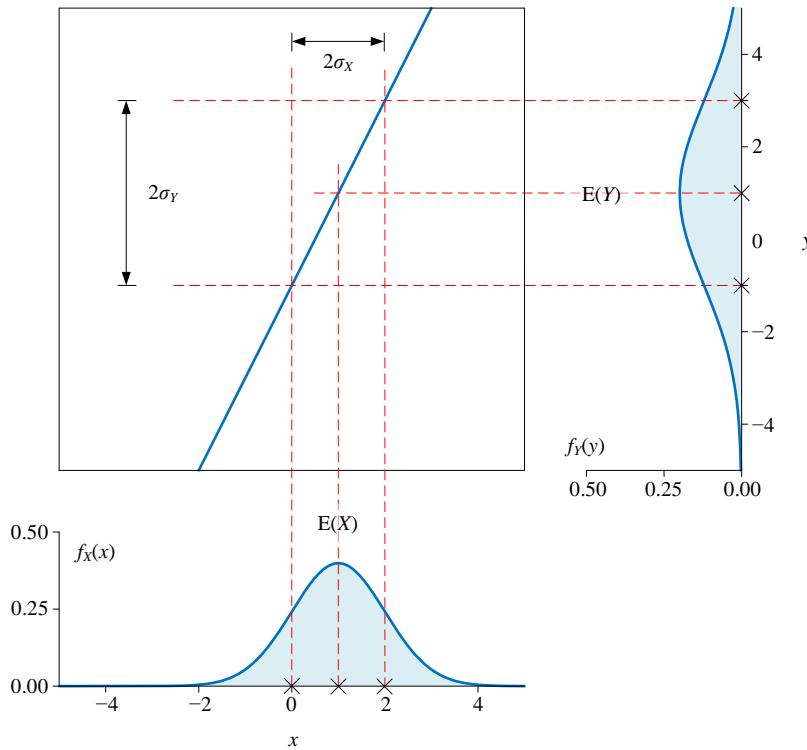


图 30. 线性变换对均值和方差的影响

面积归 1

$f_X(x)$ 作为一个一元随机变量的概率密度函数的基本要求：1) 非负；2) 面积为 1：

$$\begin{aligned} f_X(x) &\geq 0 \\ \int_{-\infty}^{+\infty} f_X(x) dx &= 1 \end{aligned} \tag{17}$$

这便解释了为什么 (1) 分母上要除以 $\sqrt{2\pi}$ ？因为如下高斯函数积分结果为 $\sqrt{2\pi}$ ：

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{2\pi} \tag{18}$$

也就是说：

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1 \tag{19}$$

下面，利用积分证明 (1) 和整个横轴围成的面积为 1：

$$\begin{aligned}
 \int_{-\infty}^{+\infty} f_x(x) dx &= \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \\
 &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) d\left(\frac{x-\mu}{\sigma}\right) \\
 &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz = \frac{\sqrt{2\pi}}{\sqrt{2\pi}} = 1
 \end{aligned} \tag{20}$$

换个角度来看，为了让把 $g(x) = \exp\left(-\frac{x^2}{2}\right)$ 改造成一个连续随机变量的 PDF，我们需要一个系数让将曲线和横轴围成的面积为 1。这个系数就是 $\frac{1}{\sqrt{2\pi}}$ ！

历史上，以下两个函数都曾多作为正态函数 PDF 解析式：

$$\begin{aligned}
 f_1(x) &= \frac{1}{\sqrt{\pi}} \exp(-x^2) \\
 f_2(x) &= \exp(-\pi x^2)
 \end{aligned} \tag{21}$$

它们之所以被大家放弃，都是因为方差不方便。 $f_1(x)$ 的方差为 $1/2$ 。 $f_2(x)$ 的方差为 $1/(2\pi)$ 。显而易见，作为标准正态分布的 PDF，(4) 更方便，因为它的方差为 1，标准差也是 1。

距离 → 亲密度

大家可能还有印象，我们在《数学要素》第 12 章讲过讲高斯函数：

$$f(x) = \exp(-x^2) \tag{22}$$

(22) 的积分为：

$$\int_{-\infty}^{\infty} \exp(-x^2) dx = \sqrt{\pi} \tag{23}$$

前文提过几次，Z 分数代表“距离”，而利用类似 (22) 这种高斯函数，我们将“距离”转换成“亲密度”。这样我们更容易理解 (10)，距离期望值 μ 越近，亲近度越大，代表可能性越大，概率密度越大；反之，离 μ 越远，越疏远，代表可能性越小，概率密度越小。本书后文还会用这个视角分析其他高斯分布。



在实际应用中，高斯分布经常用于建模和分析连续型数据，如测量值、物理量和经济指标等。在机器学习和数据分析中，高斯分布也被广泛应用于分类、聚类、离群点检测等问题中。但是，仅仅掌握一元高斯分布的知识是不够的。从下一章开始，我们将探讨二元、多元高斯分布、条件高斯分布，以及高斯分布背后的协方差矩阵。

10

Bivariate Gaussian Distribution

二元高斯分布

椭圆的影子几乎无处不在



自然之书是用数学语言写成的，符号是三角形、圆形和其他几何图形；不理解几何图形，别想读懂自然之书；没有它们，我们只能在黑暗的迷宫中徘徊不前。

The book of nature is written in mathematical language, and the symbols are triangles, circles and other geometrical figures, without whose help it is impossible to comprehend a single word of it; without which one wanders in vain through a dark labyrinth.

——伽利略·伽利莱 (Galilei Galileo) | 意大利物理学家、数学家及哲学家 | 1564 ~ 1642



- ◀ `matplotlib.patches.Rectangle()` 绘制长方形
- ◀ `matplotlib.pyplot.axhline()` 绘制水平线
- ◀ `matplotlib.pyplot.axvline()` 绘制竖直线
- ◀ `matplotlib.pyplot.contour()` 绘制等高线图
- ◀ `matplotlib.pyplot.contourf()` 绘制填充等高线图
- ◀ `scipy.stats.multivariate_normal()` 多元高斯分布
- ◀ `scipy.stats.multivariate_normal.cdf()` 多元高斯分布 CDF 函数
- ◀ `scipy.stats.multivariate_normal.pdf()` 多元高斯分布 PDF 函数

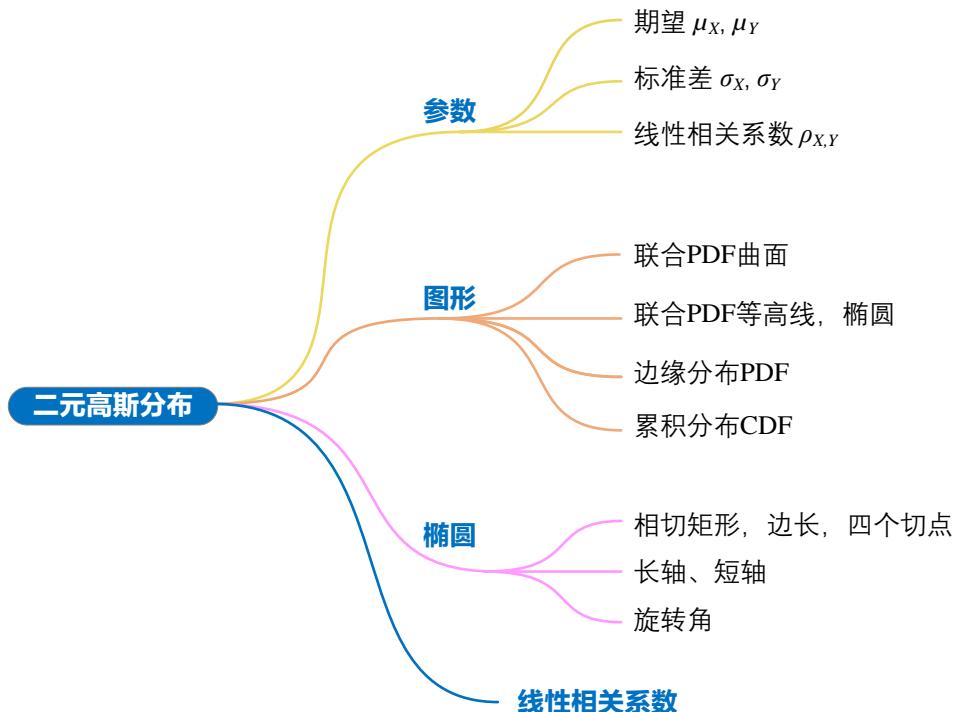
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



10.1 二元高斯分布：看见椭圆

概率密度函数

二元高斯分布 (bivariate Gaussian distribution), 也叫**二元正态分布** (bivariate normal distribution), 它的概率密度函数 $f_{X,Y}(x,y)$ 解析式如下：

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \times \exp\left(\underbrace{\frac{-1}{2}\frac{1}{(1-\rho_{X,Y}^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)}_{\text{Ellipse}}\right) \quad (1)$$

其中, μ_X 和 μ_Y 分别为随机变量 X 、 Y 的期望值, σ_X 和 σ_Y 分别为随机变量 X 、 Y 的标准差, $\rho_{X,Y}$ 为 X 和 Y 线性相关系数。分母中, 系数 $2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}$ 完成归一化, 也就是让 $f_{X,Y}(x,y)$ 和水平面围成的体积为 1。

注意, 观察 (1), 显然 $\rho_{X,Y}$ 取值区间为 $(-1, 1)$, 不能为 ± 1 ; 否则, 分母为 0。

此外, 丛书之前反复提到二元高斯分布和椭圆的关系。我们在 (1) 中已经看到了椭圆解析式。



(1) 中蕴含的椭圆解析式形式正是我们在《数学要素》第 9 章讲过的特殊类型。

PDF 曲面形状

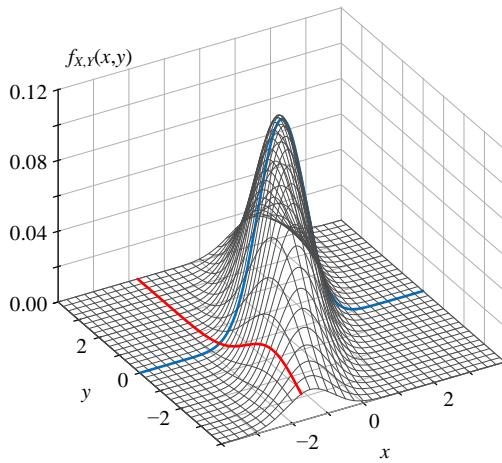
给定如下条件:

$$\mu_X = 0, \quad \mu_Y = 0, \quad \sigma_X = 1, \quad \sigma_Y = 2, \quad \rho_{X,Y} = 0.75 \quad (2)$$

绘制满足条件的二元正态分布密度函数曲面, 具体如图 1 所示。

容易发现, μ_X 和 μ_Y 决定曲面中心所在位置; σ_X 和 σ_Y 影响曲面在 x 和 y 方向的形状。而 $\rho_{X,Y}$ 似乎提供了曲面的扭曲。实际上, σ_X 、 σ_Y 、 $\rho_{X,Y}$ 都影响了曲面的倾斜。

下面, 我们从几个侧面来深入观察二元高斯分布 $PDF f_{X,Y}(x,y)$ 曲面。

图 1. 二元高斯分布 PDF 函数曲面 $f_{X,Y}(x,y)$, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

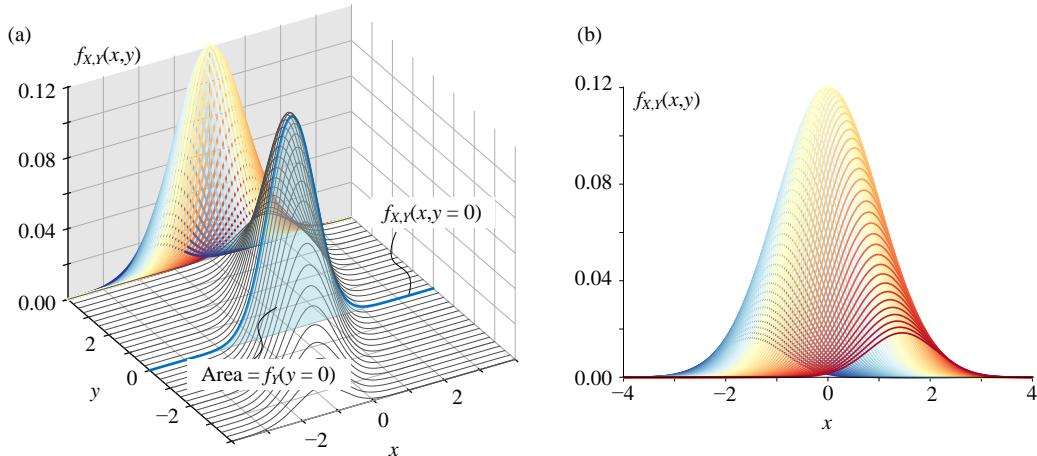
沿 x 剖面线

图 2 所示为 $f_{X,Y}(x,y)$ 曲面沿 x 方向的剖面线，以及这些曲线在 xz 平面上的投影。这些曲线，相当于是(1)中 y 取定值时 PDF 对应的曲线。比如 $y=0$ 时，曲线的解析式：

$$f_{X,Y}(x, y=0) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{X,Y}^2}} \times \exp\left(\frac{-1}{2}\frac{1}{(1-\rho_{X,Y}^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \frac{2\rho_{X,Y}\mu_Y}{\sigma_Y}\left(\frac{x-\mu_X}{\sigma_X}\right) + \left(\frac{\mu_Y}{\sigma_Y}\right)^2\right)\right) \quad (3)$$

观察这条曲线，我们都能看到一元正态分布的影子。

注意，举个例子，图 2 (a) 中 $f_{X,Y}(x,y=0)$ 这条曲线和横轴围成的图形面积并不为 1，面积对应边缘 PDF $f_Y(y=0)$ 。因此图 2 (b) 中这些曲线虽然看起来像一元高斯分布 PDF，实际上并不是。但是经过一定的缩放，它们可以成为条件高斯分布的 PDF。

图 2. PDF 函数曲面 $f_{X,Y}(x,y)$, 沿 x 方向的剖面线, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

大家试想一下，如果我们可以得到 $y = 0$ 时边缘 PDF $f_Y(y=0)$ 的具体值，就可以利用贝叶斯定理得到条件概率 $f_{X|Y}(x | y=0)$ ：

$$f_{X|Y}(x | y=0) = \frac{f_{X,Y}(x, y=0)}{f_Y(y=0)} \quad (4)$$

分母中的 $f_Y(y=0)$ 起到归一化的作用。而 $f_{X|Y}(x | y=0)$ 摆身一变成了条件高斯分布的 PDF。



这是本书第 12 章要讲解的内容。

沿 y 剖面线

图 3 所示为 $f_{X,Y}(x,y)$ 曲面沿 y 方向的剖面线，以及这些曲线在 yz 平面上的投影。曲线相当于 x 取定值，联合 PDF $f_{X,Y}(x,y)$ 随 y 变化。

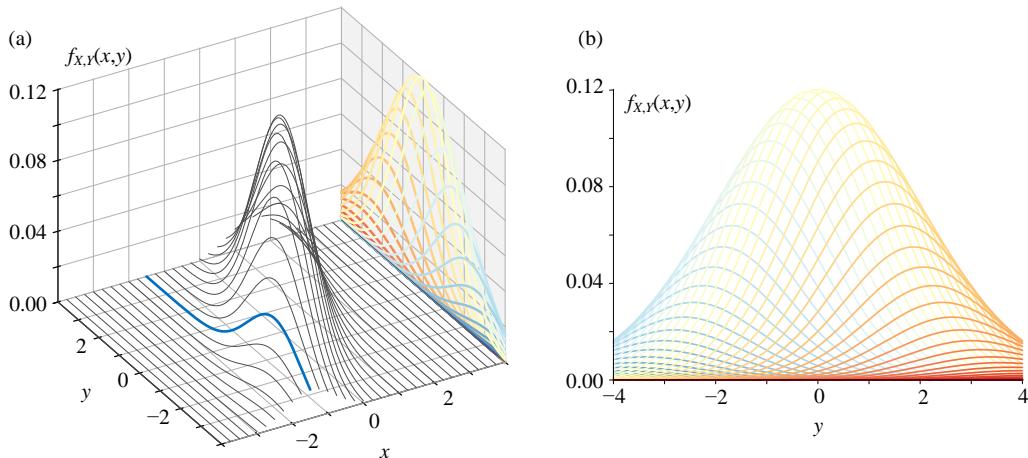
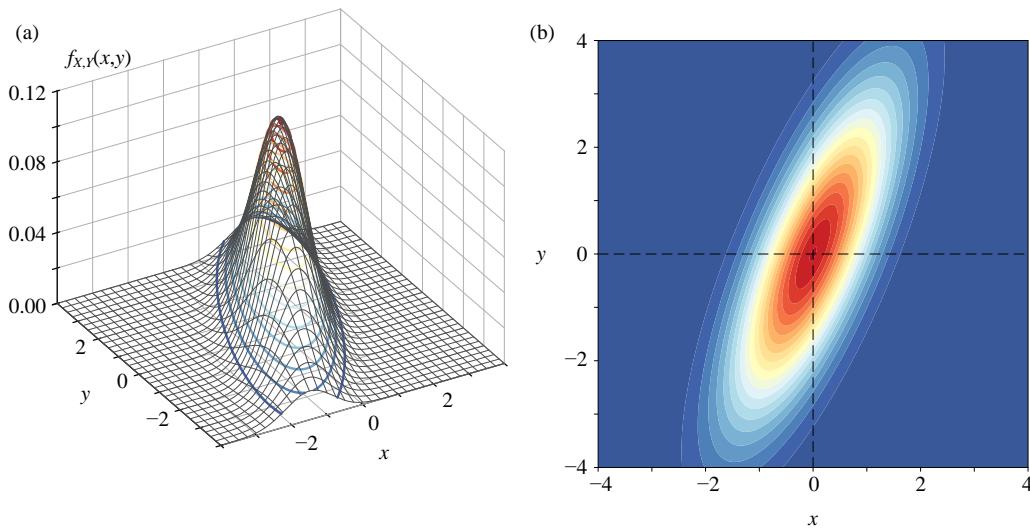


图 3. PDF 函数曲面 $f_{X,Y}(x,y)$, 沿 y 方向的剖面线, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

等高线

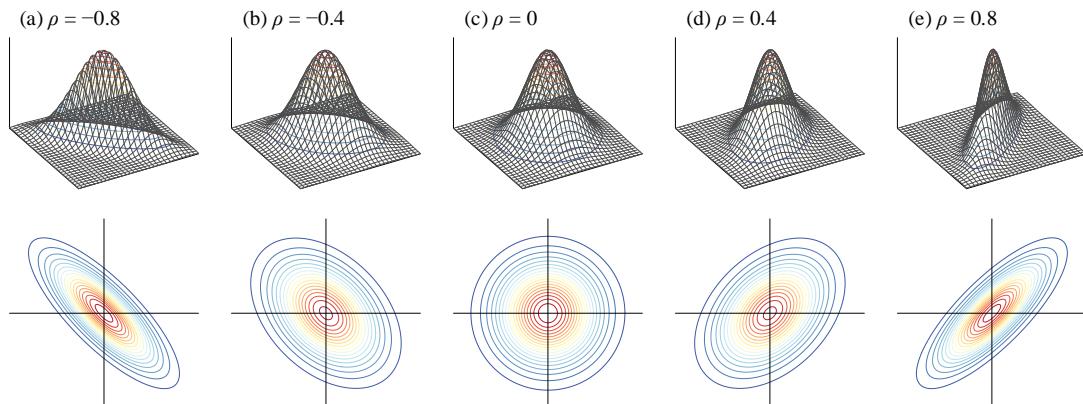
图 4 所示为 $f_{X,Y}(x,y)$ 曲面等高线。很明显，我们已经从等高线中看到椭圆。特别是在图 4 (b) 中，我们看到一系列同心旋转椭圆。这并不奇怪，因为 (1) 中 $\exp()$ 函数中蕴含着一个椭圆解析式。

这也就是为什么高斯分布被称作是一种**椭圆分布** (elliptical distribution)。本章后续将揭开高斯分布和椭圆的更多联系。

图 4. PDF 函数曲面 $f_{x,y}(x,y)$, 空间等高线和平面填充等高线, $\sigma_x = 1, \sigma_y = 2, \rho_{x,y} = 0.75$

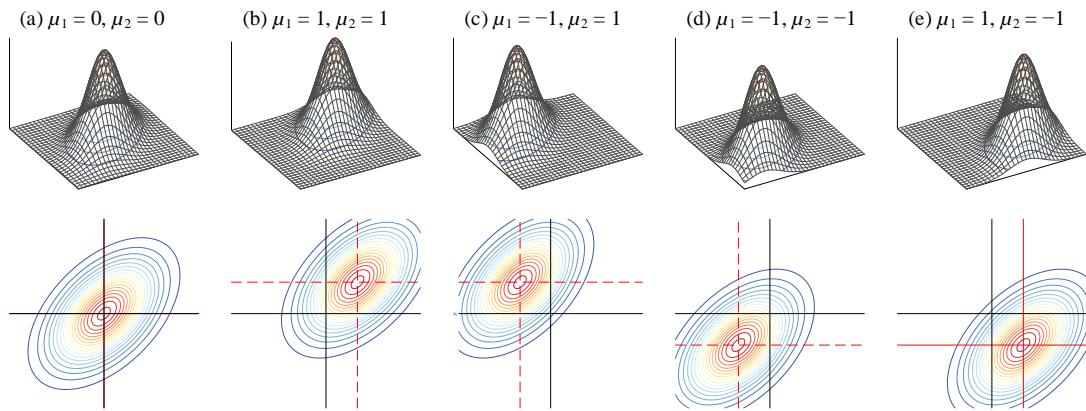
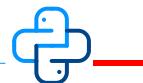
相关性系数

为了方便大家了解相关性系数对二元高斯分布 PDF 的影响, 设定 $\sigma_x = 1, \sigma_y = 1$ 。如图 5 所示, 相关性系数对二元高斯分布 PDF 曲面和等高线形状的影响。

图 5. 不同相关性系数, 二元高斯分布 PDF 曲面和等高线, $\sigma_x = 1, \sigma_y = 1$

质心

如图 6 所示, 固定相关性系数和标准差, 改变质心仅仅影响曲面中心位置。

图 6. 不同质心位置，二元高斯分布 PDF 曲面和等高线， $\sigma_x = 1, \sigma_y = 1$ 

Bk5_Ch10_01.py 绘制本节图像。



在 Bk5_Ch10_01.py 基础上，我们用 Streamlit 制作了一个应用，大家可以改变 $\rho_{X,Y}$ 、 σ_X 、 σ_Y 三个参数，观察二元高斯 PDF 曲面、等高线变化。请大家参考 Streamlit_Bk5_Ch10_01.py。

10.2 边缘分布：一元高斯分布

边缘分布

大家可能也已经注意到，不考虑 Y 的时候， X 应该服从一元高斯分布。而 μ_X 和 σ_X 是描述随机变量 X 的参数。也就是说，有了这两个参数，我们就可以写出 X 的边缘 PDF $f_X(x)$ —— 一元高斯分布概率密度函数：

$$f_X(x) = \frac{1}{\sigma_X \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - \mu_X}{\sigma_X}\right)^2\right) \quad (5)$$

同理， μ_Y 和 σ_Y 是描述随机变量 Y 的参数，对应写出 Y 的边缘 PDF $f_Y(y)$ ：

$$f_Y(y) = \frac{1}{\sigma_Y \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - \mu_Y}{\sigma_Y}\right)^2\right) \quad (6)$$

在图4平面等高线基础上添加 $f_X(x)$ 和 $f_Y(y)$ 边缘 PDF 图像子图，我们便得到图7。

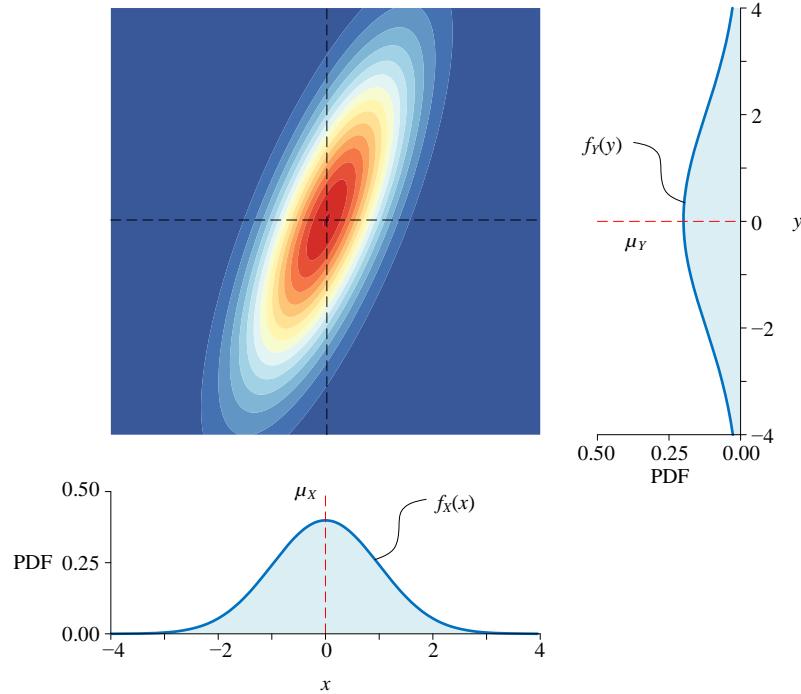


图 7. 二元高斯分布 PDF 和边缘 PDF, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

偏积分求边缘分布 PDF

下面，以 Y 的边缘分布概率密度函数 $f_Y(y)$ 为例证明二元高斯分布 PDF“偏积分”得到一元高斯分布 PDF。

连续随机变量 Y 的边缘分布概率密度函数 $f_Y(y)$ 可以通过 $f_{X,Y}(x,y)$ 对 x 偏积分得到，即：

$$f_Y(y) = \overbrace{\int_{-\infty}^{+\infty} f_{X,Y}(x,y) dx}^{\text{Eliminate } x} \quad (7)$$

令，

$$G(x,y) = \frac{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}{(1-\rho_{X,Y}^2)} \quad (8)$$

这样，二元高斯分布可以写成：

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \times \exp\left(-\frac{1}{2}G(x,y)\right) \quad (9)$$

将(8)中 $G(x,y)$ 写成：

$$\begin{aligned} G(x,y) &= \frac{\left(\frac{x-\mu_x}{\sigma_x} - \rho_{x,y} \frac{y-\mu_y}{\sigma_y}\right)^2}{\left(1-\rho_{x,y}^2\right)} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \\ &= \frac{\left(x - \left(\mu_x + \rho_{x,y} \frac{\sigma_x}{\sigma_y} (y - \mu_y)\right)\right)^2}{\left(1-\rho_{x,y}^2\right) \sigma_x^2} + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \end{aligned} \quad (10)$$

令

$$t = t(y) = \mu_x + \rho_{x,y} \frac{\sigma_x}{\sigma_y} (y - \mu_y) \quad (11)$$

可以发现 t 仅仅是 y 的函数，与 x 无关，这样便于积分。

$G(x,y)$ 进一步整理为：

$$G(x,y) = \frac{(x-t)^2}{\left(1-\rho_{x,y}^2\right) \sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} \quad (12)$$

将(12)代入(9)得到：

$$f_{x,y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \times \exp\left(-\frac{1}{2}\left(\frac{(x-t)^2}{\left(1-\rho_{x,y}^2\right)\sigma_x^2}\right)\right) \times \exp\left(-\frac{1}{2}\left(\frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right) \quad (13)$$

将(13)代入(7)得到：

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{+\infty} \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \times \exp\left(-\frac{1}{2}\left(\frac{(x-t)^2}{\left(1-\rho_{x,y}^2\right)\sigma_x^2}\right)\right) \times \exp\left(-\frac{1}{2}\left(\frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right) dx \\ &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right) \cdot \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}\left(\frac{(x-t)^2}{\left(\sqrt{\left(1-\rho_{x,y}^2\right)\sigma_x^2}\right)^2}\right)\right) dx \end{aligned} \quad (14)$$

回忆，我们在《数学要素》讲解过高斯函数积分：

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2}\left(\frac{(x-t)^2}{\left(\sqrt{\left(1-\rho_{x,y}^2\right)\sigma_x^2}\right)^2}\right)\right) dx = \sqrt{2\pi} \sqrt{1-\rho_{x,y}^2} \sigma_x \quad (15)$$

将(15)代入(14)，得到：

$$\begin{aligned}
 f_Y(y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}} \cdot \exp\left(\frac{-1}{2}\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right) \sqrt{2\pi} \sqrt{(1-\rho_{X,Y}^2)}\sigma_X \\
 &= \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(\frac{-1}{2}\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)
 \end{aligned} \tag{16}$$

⚠ 再次强调，联合 PDF $f_{X,Y}(x,y)$ 二重积分得到的是概率，也就是曲面体积代表概率。而 $f_{X,Y}(x,y)$ 偏积分得到的还是概率密度，即边缘概率密度 $f_X(x)$ 或 $f_Y(y)$ 。边缘 PDF $f_X(x)$ 和 $f_Y(y)$ 进一步积分才得到概率。

独立

图 8 所示为二元高斯分布参数对 PDF 等高线影响。

特别地，当相关性系数 $\rho_{X,Y}$ 为 0 时：

$$\begin{aligned}
 f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y} \times \exp\left(\frac{-1}{2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right) \\
 &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(\frac{-1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \times \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left(\frac{-1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right) \\
 &= f_X(x)f_Y(y)
 \end{aligned} \tag{17}$$

观察图 8 (b)、(e)、(h)，我们发现椭圆等高线为正椭圆。

⚠ 注意，独立意味着两个变量的取值之间没有任何关系，即它们的联合概率分布等于它们的边缘概率分布的乘积。而相关则表示两个变量之间存在某种形式的关联关系，可以是线性的，也可以是非线性的。因此，线性相关系数为 0 只是说明两个变量之间不存在线性关系，但并不能推断它们是否独立。

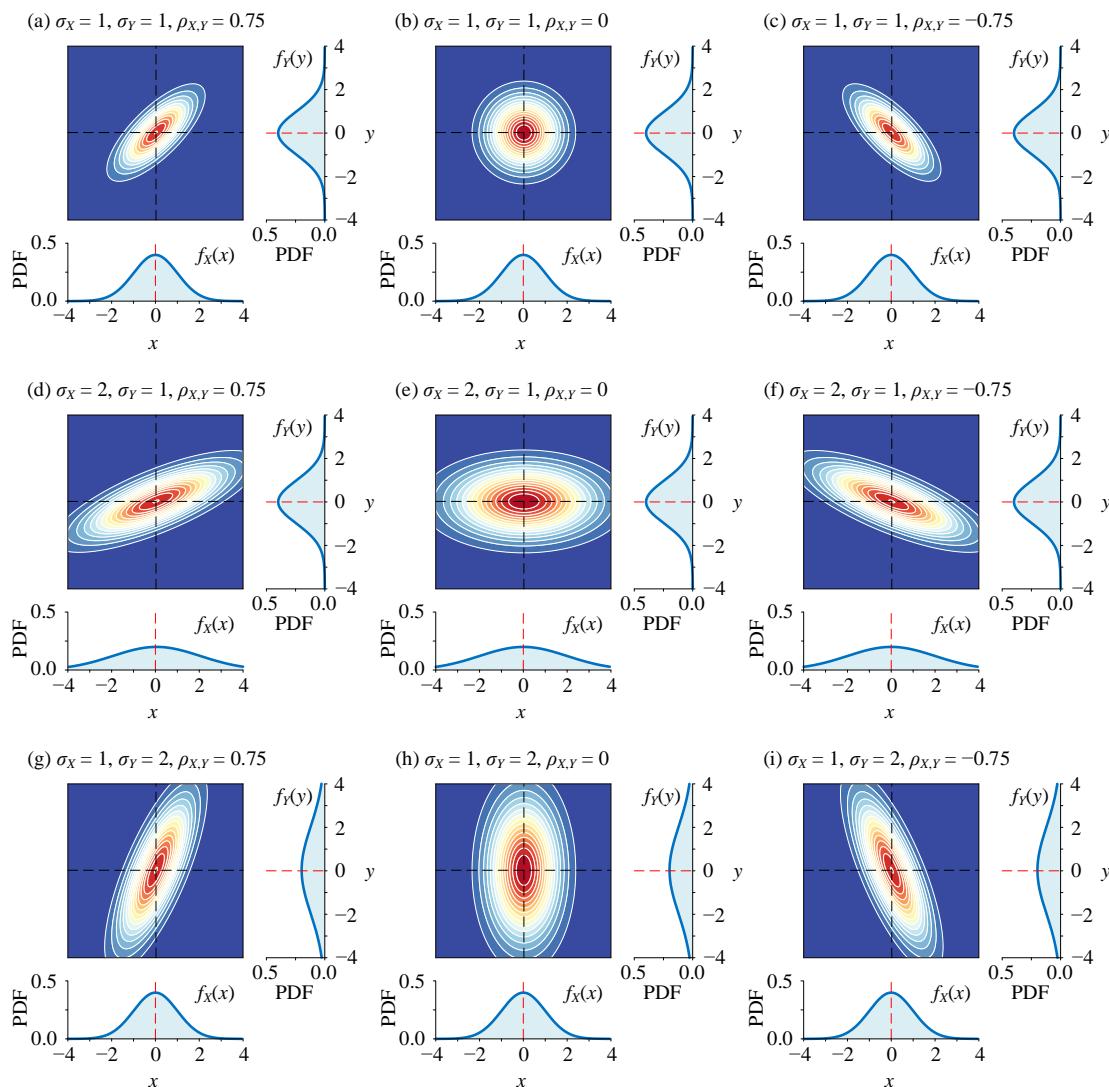


图 8. 二元高斯分布参数对 PDF 等高线影响



Bk5_Ch10_02.py 绘制本节图像。请大家自行调整分布参数。

10.3 累积分布函数：概率值

二元高斯分布的累积分布函数 $CDF F_{X,Y}(x,y)$ 是对 PDF $f_{X,Y}(x,y)$ 的二重积分：

$$F_{X,Y}(x,y) = \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(s,t) ds dt \quad (18)$$

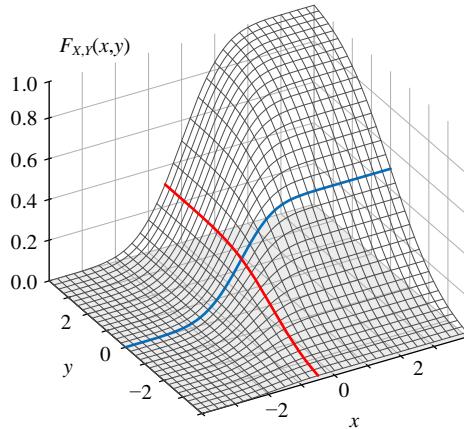
图 9 所示为二元高斯分布累积分布函数 CDF 曲面。

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

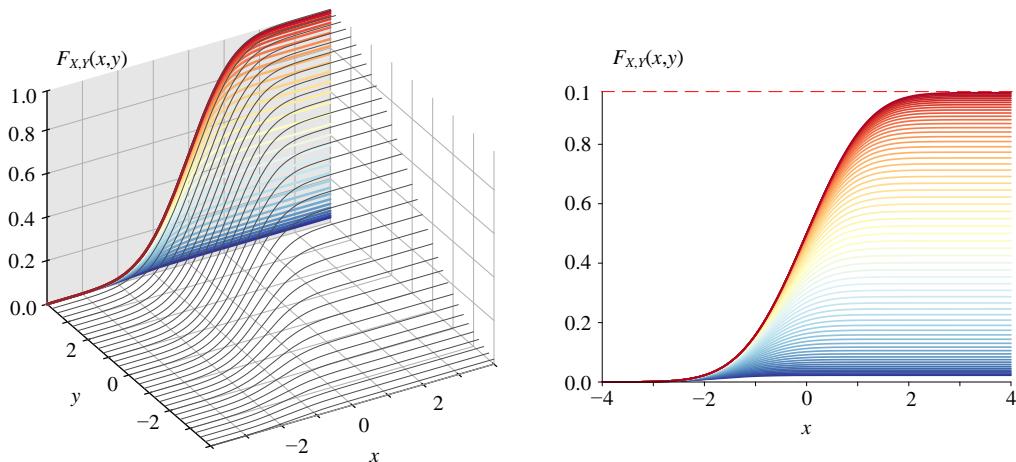
本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 9. 二元高斯分布累积函数 CDF 曲面, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

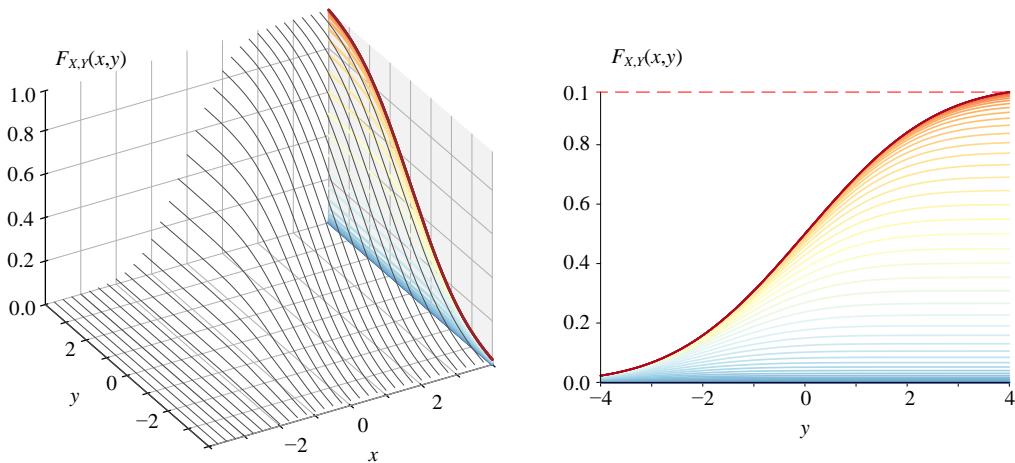
沿 x 剖面线

和上一节一样，下面从几个侧面来观察二元高斯分布 CDF 曲面 $F_{X,Y}(x,y)$ 。图 10 所示为 $F_{X,Y}(x,y)$ 曲面沿 x 方向的剖面线，以及这些曲线在 xz 平面上的投影。

图 10. CDF 曲面 $F_{X,Y}(x,y)$, 沿 x 方向的剖面线, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

沿 y 剖面线

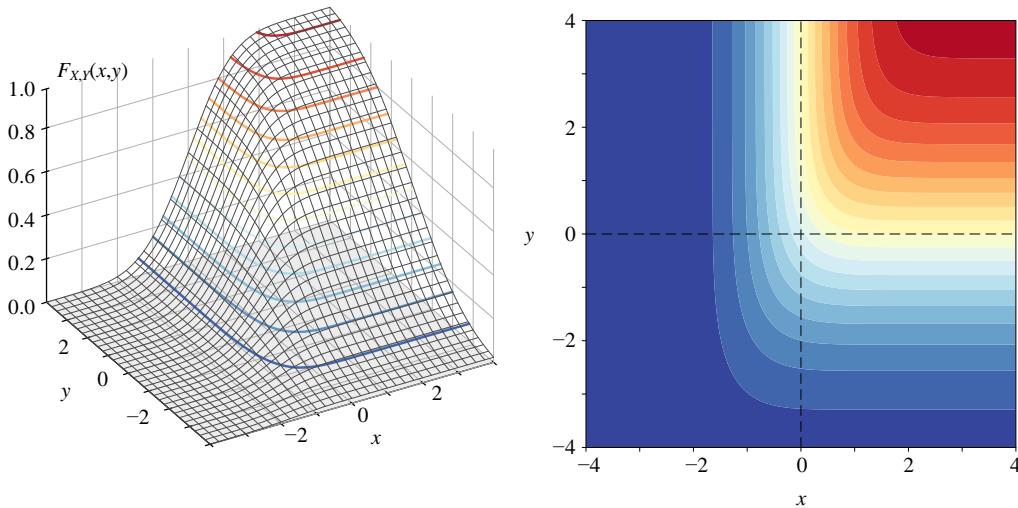
图 11 所示为 $F_{X,Y}(x,y)$ 曲面沿 y 方向的剖面线，以及这些曲线在 yz 平面上的投影。

图 11. CDF 曲面 $F_{X,Y}(x,y)$, 沿 y 方向的剖面线, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

等高线

图 12 所示为 CDF 函数曲面 $F_{X,Y}(x,y)$ 的等高线。图 13 所示为，在 $F_{X,Y}(x,y)$ 的平面填充等高线基础上，又绘制了边缘 CDF $F_X(x)$ 、 $F_Y(y)$ 曲线。

请大家修改上一节代码绘制本节图像。只需要把 `scipy.stats.multivariate_normal.pdf()` 换成 `scipy.stats.multivariate_normal.cdf()` 函数。

图 12. CDF 函数曲面 $F_{X,Y}(x,y)$, 空间等高线和平面填充等高线, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

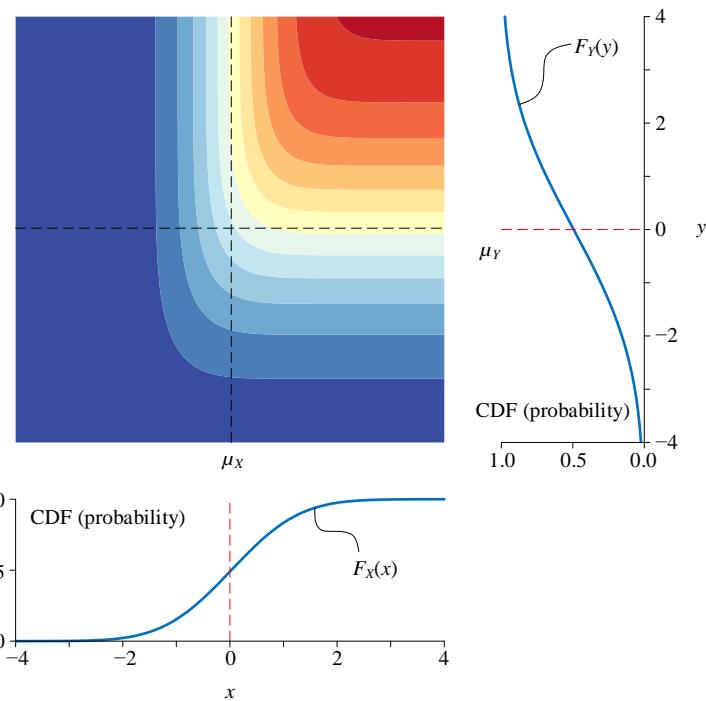
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 13. CDF 函数曲面 $F_{X,Y}(x,y)$ 平面填充等高线，边缘概率分布 CDF

10.4 用椭圆解剖二元高斯分布

大家已经在 (1) 看到了椭圆的解析式，这一节我们对二元高斯分布和椭圆的关系进行定量研究。

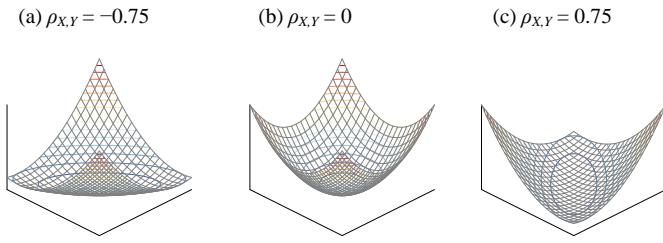
二次曲面

利用 (8) 中定义的 $G(x,y)$ 。将 (8) 代入 (1)，得到：

$$f_{x,y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \times \exp\left(\frac{-1}{2}G(x,y)\right) \quad (19)$$

图 14 所示为 $G(x,y)$ 代表的几种曲面。

但是，对于二元高斯分布来说，如果 PDF 解析式存在，相关性的取值范围为 $(-1, 1)$ ，此时协方差矩阵为正定。请大家思考如果，协方差矩阵为半正定， $G(x,y)$ 曲面的形状是什么？

图 14. $G(x, y)$ 代表的几种曲面

椭圆

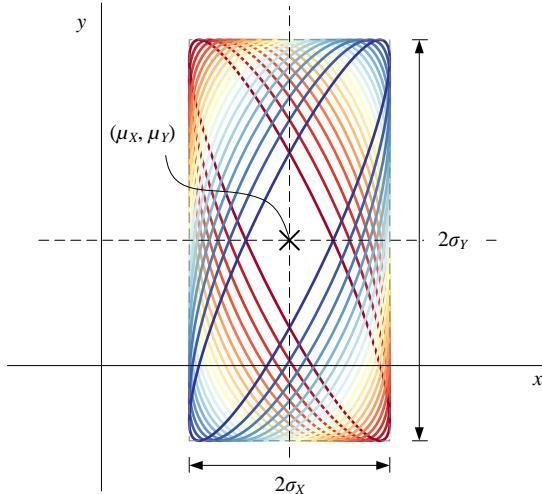
令 $G(x, y) = 1$, 当 $\rho_{X,Y}$ 在 $(-1, 1)$ 变化时, 我们便得到椭圆的解析式:

$$\frac{1}{(1-\rho_{X,Y}^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho_{X,Y} \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] = 1 \quad (20)$$

(μ_1, μ_2) 确定椭圆中心位置, σ_1 、 σ_2 和 ρ 三者共同决定椭圆长短轴长度和旋转角度。

“鸢尾花书”《数学要素》第 9 章介绍过, 形如 (20) 解析式的椭圆有重要的特点——椭圆和长 $2\sigma_X$ 、宽 $2\sigma_Y$ 的矩形相切。

图 15 给出的矩形框中心位于 (μ_X, μ_Y) , 矩形框长度为 $2\sigma_X = 2$ 、宽度为 $2\sigma_Y = 4$ 。图 15 中一系列椭圆对应的相关性系数 $\rho_{X,Y}$ 的变化范围为 $[-0.9, 0.9]$ 。

图 15. 椭圆和中心在 (μ_X, μ_Y) 长 $2\sigma_X$ 、宽 $2\sigma_Y$ 的矩形相切

相关性系数 ρ 大于 0 时, 即线性正相关, 椭圆长轴指向约东北方向。线性相关性系数 ρ 小于 0 时, 即负相关, 椭圆长轴指向约西北方向; 特别提醒读者注意的是, 当相关性系数 ρ 为 0 时, 椭圆为正椭圆。

图 16 所示为三种标准差 σ_X 、 σ_Y 大小不同的情况，和矩形相切的椭圆随着相关性系数变化情况。

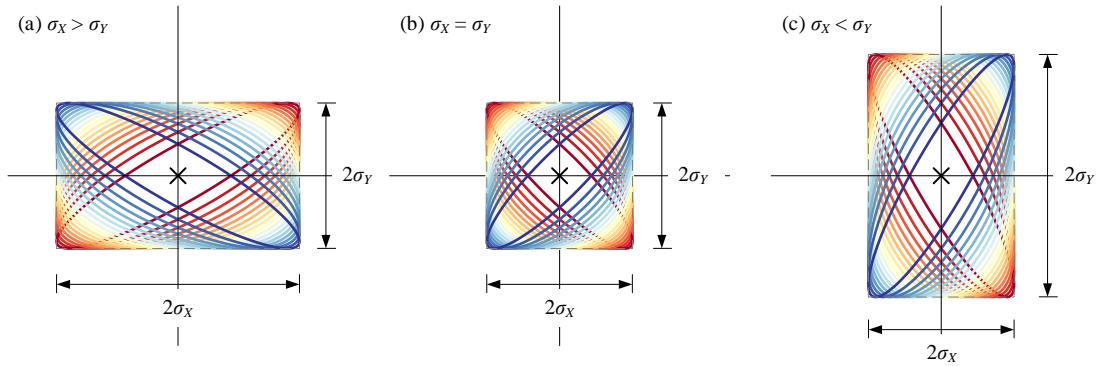


图 16. 三种标准差 σ_X 、 σ_Y 大小不同的情况

四个切点

椭圆和矩形有四个切点，下面我们来求解这四个切点的具体位置。考虑特殊情况 $\mu_X=0, \mu_Y=0$, (20) 可以简化为：

$$\frac{1}{(1-\rho_{X,Y}^2)} \left(\left(\frac{x}{\sigma_X} \right)^2 - \frac{2\rho_{X,Y}}{\sigma_X \sigma_Y} xy + \left(\frac{y}{\sigma_Y} \right)^2 \right) = 1 \quad (21)$$

将 $y = \sigma_Y$ 代入 (21)，得到：

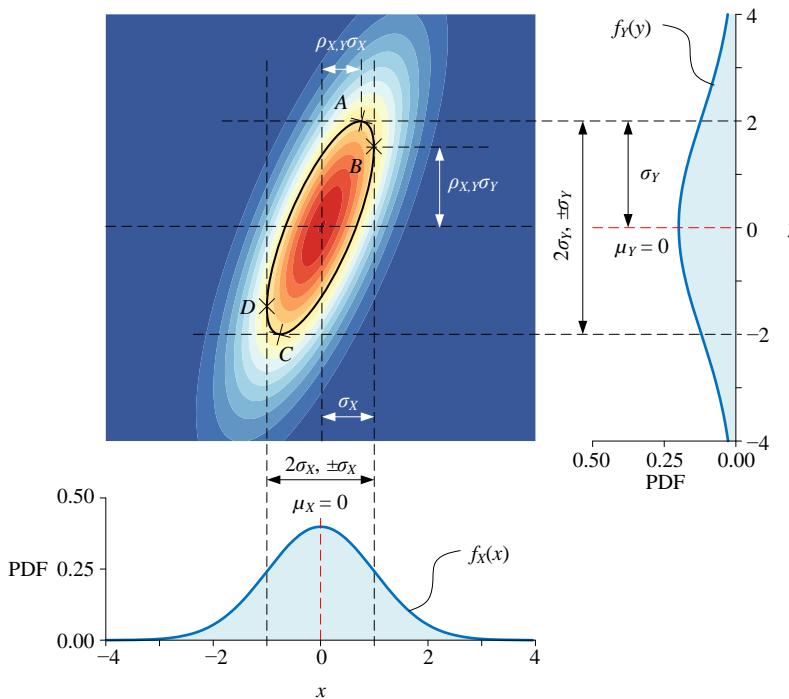
$$\left(\frac{x}{\sigma_X} - \rho_{X,Y} \right)^2 = 0 \quad (22)$$

这样我们便得到一个切点：

$$\begin{cases} x = \rho_{X,Y} \sigma_X \\ y = \sigma_Y \end{cases} \quad (23)$$

同理，获得所有四个切点 A、B、C、D 的具体位置：

$$A(\rho_{X,Y} \sigma_X, \sigma_Y), \quad B(\sigma_X, \rho_{X,Y} \sigma_Y), \quad C(-\rho_{X,Y} \sigma_X, -\sigma_Y), \quad D(-\sigma_X, -\rho_{X,Y} \sigma_Y) \quad (24)$$

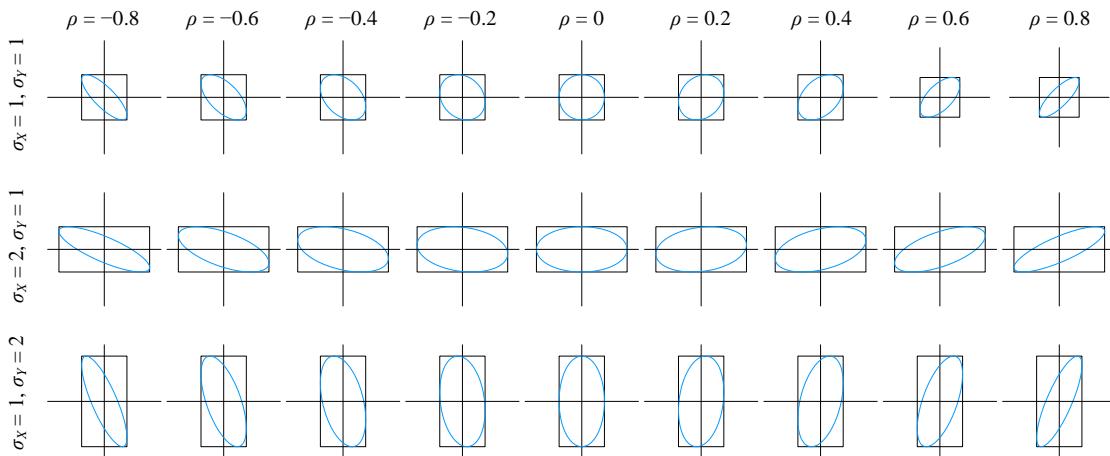
图 17. 二元高斯分布 PDF 和边缘 PDF, $\sigma_X = 1, \sigma_Y = 2, \rho_{X,Y} = 0.75$

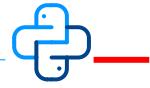
椭圆和矩形

μ_X 和 μ_Y 均不为 0 的一般情况，四个切点的位置平移 (μ_X, μ_Y) ，为：

$$\begin{aligned} &A(\mu_X + \rho_{X,Y}\sigma_X, \mu_Y + \sigma_Y), \quad B(\mu_X + \sigma_X, \mu_Y + \rho_{X,Y}\sigma_Y), \\ &C(\mu_X - \rho_{X,Y}\sigma_X, \mu_Y - \sigma_Y), \quad D(\mu_X - \sigma_X, \mu_Y - \rho_{X,Y}\sigma_Y) \end{aligned} \quad (25)$$

图 18 所示为椭圆和矩形切点位置随 σ_X 、 σ_Y 、 $\rho_{X,Y}$ 变化关系，请大家自行总结规律。

图 18. 椭圆和矩形切点随 σ_X 、 σ_Y 、 $\rho_{X,Y}$ 变化关系



Bk5_Ch10_03.py 绘制图 18。

椭圆形状

再怎么强调椭圆和高斯分布的紧密联系也不为过。图 19 这个旋转椭圆的位置、形状、旋转角度等信息，蕴含着高斯分布的中心 (μ_x, μ_y) 、标准差 σ_x 和 σ_y 、相关性系数 $\rho_{x,y}$ 。也就是说，某个二元高斯分布可以用特定椭圆来代表。

图 19 中还有很多椭圆相关的性质值得我们挖掘。

图 19 所示两个椭圆，蓝色椭圆上所有点代入 (8) 都等于 1，类似一元高斯分布中的 $\mu \pm \sigma$ 。而更大一点的红色椭圆所有点代入 (8) 都等于 4，平方根为 2，类似一元高斯分布中的 $\mu \pm 2\sigma$ 。

上面所述的平方根 (1, 2) 正是《矩阵力量》第 20 章讲过马氏距离。本章后文将稍微回顾马氏距离，本书第 23 章还要深入讲解马氏距离。

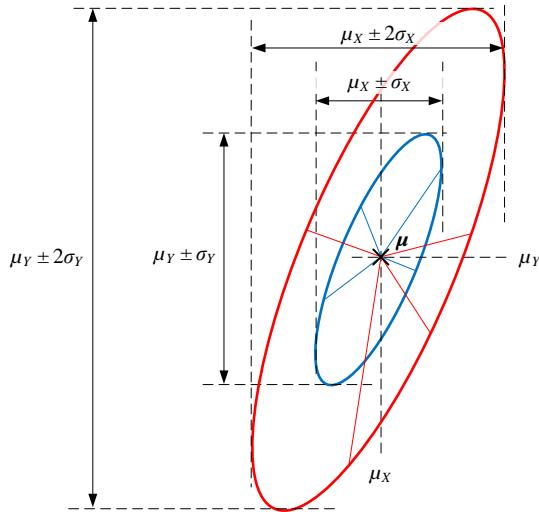


图 19. 两个椭圆

图 20 中的浅蓝色直角三角形的两条直角边长度分别是 $\rho_{x,y}\sigma_y$ 、 σ_x ，其中 θ 角的正切值为：

$$\tan \theta = \frac{\rho_{x,y}\sigma_y}{\sigma_x} \quad (26)$$

图 20 所示 AC 线段、 BD 线段和条件概率、线性回归有着直接联系。本书第 12 章将专门讲解高斯分布条件概率。



图 20 中两条红色线为椭圆的长轴和短轴所在方向，这两条直线又和主成分分析有着密切的关系。这是本书第 14、25 章要探讨的内容。

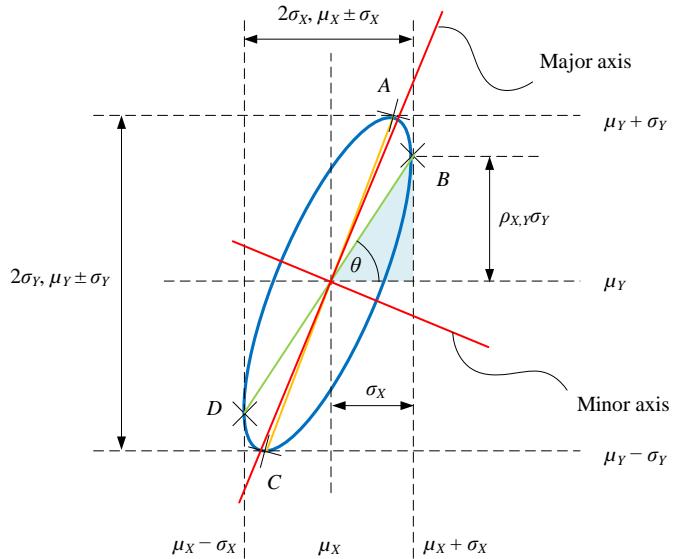


图 20. 椭圆中的四条直线

10.5 聊聊线性相关性系数

几种可视化方案

图 21 所示为相关性系数的几种可视化方案，比如散点图、二元高斯 PDF 曲面、PDF 等高线、条件概率直线、向量夹角。



大家应该在《矩阵力量》第 23 章见过图 21，当时我们特别讨论了利用向量可视化线性相关系数。

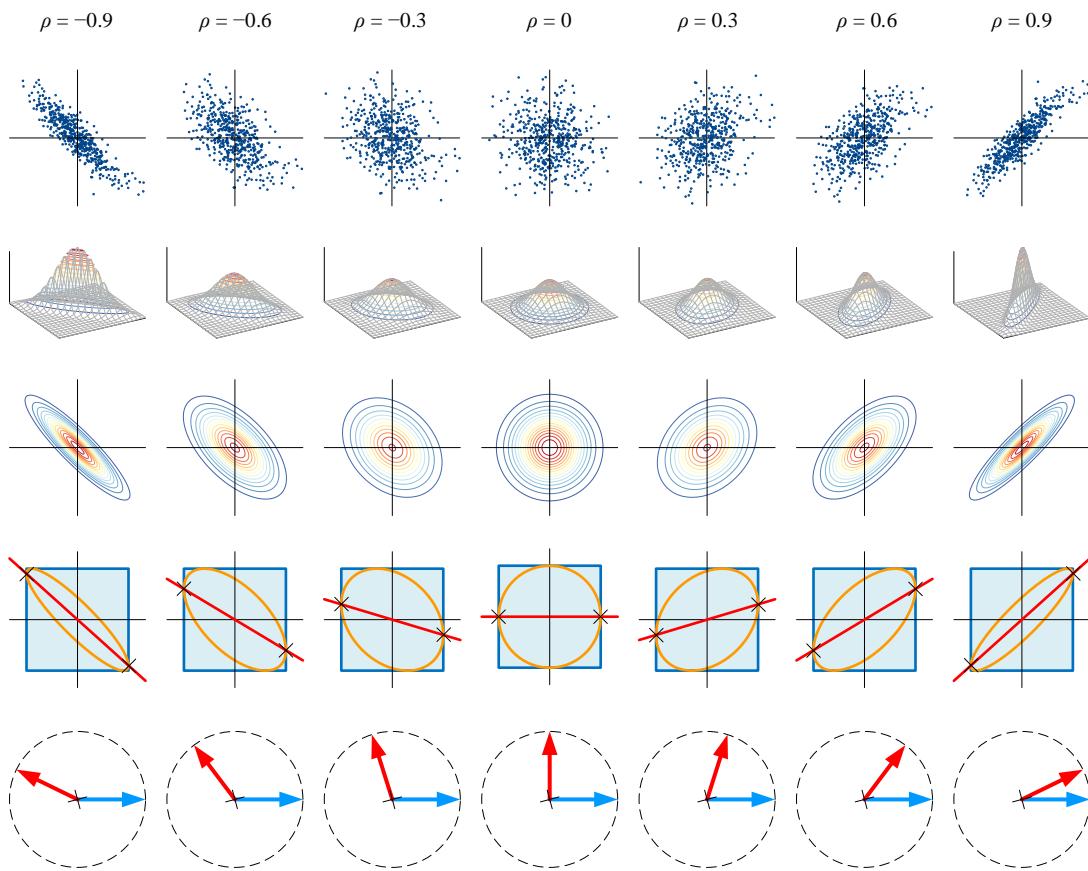


图 21. 相关性系数的几种可视化方案



Bk5_Ch10_04.py 可以绘制图 21 大部分图像。请大家自行修改参数。

独立 vs 线性相关性系数为 0

本章前文提过，线性相关系数反映的是两个随机变量间的线性关系，但是随机变量之间除了线性关系还可能存在其它关系。

举个例子，随机变量 X 在 $[-1, 1]$ 连续均匀分布。令 $Y = X^2$ ，显然， X 和 Y 存在二次关系，并不独立。但是两者的协方差为 0：

$$\begin{aligned}
 \text{cov}(X, Y) &= \text{cov}(X, X^2) \\
 &= E[X \cdot X^2] - E[X] \cdot E[X^2] \\
 &= E[X^3] - E[X]E[X^2] \\
 &= 0 - 0 \cdot E[X^2] = 0
 \end{aligned} \tag{27}$$

这意味着的线性相关性系数为 0。

安斯库姆四重奏

图 22 是 **安斯库姆四重奏** (Anscombe's quartet) 的四组散点图。观察图中四组散点图，我们可以发现数据的关系完全不同。但是，它们的相关性系数几乎完全一致。

这幅图告诉我们，线性相关性系数不是万能的，它只适合度量随机变量之间的“线性关系”。此外，线性相关性系数特别容易受到**离群值** (outlier) 的影响，这一点可以从图 22 (c) 看出来。

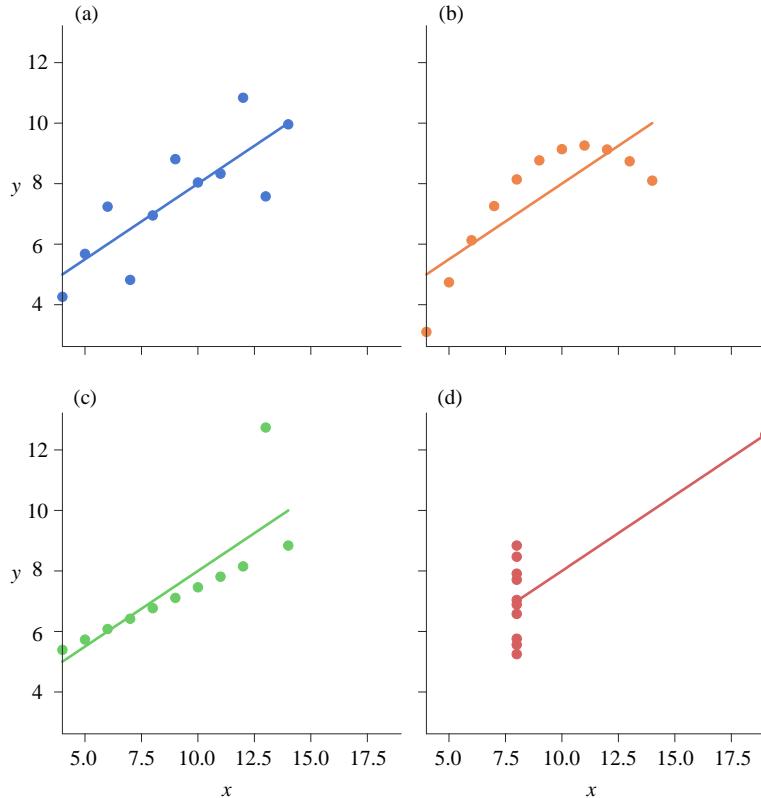


图 22. 安斯库姆四重奏

向量空间：线性无关

我们在《矩阵力量》第 7 章中介绍过，给定向量组 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_D]$ ，如果存在不全为零 $\alpha_1, \alpha_2, \dots, \alpha_D$ 使得下式成立。

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \alpha_3 \mathbf{v}_3 + \dots + \alpha_D \mathbf{v}_D = \mathbf{0} \quad (28)$$

则称向量组 \mathbf{V} **线性相关** (linear dependence); 否则， \mathbf{V} **线性无关** (linear independence)。请大家注意区分。

正交 vs 线性相关性系数为 0

随机变量 X 和 Y 的协方差可以通过下式计算得到：

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) \quad (29)$$

如果 X 和 Y 独立，则

$$\text{cov}(X, Y) = 0 \quad (30)$$

这意味着：

$$E(XY) = E(X)E(Y) \quad (31)$$

本书前文提过，随机变量 X 和 Y 的有序样本集合看做是向量 \mathbf{x} 和 \mathbf{y} 。如果向量 \mathbf{x} 和 \mathbf{y} 内积为 0，这意味着 \mathbf{x} 和 \mathbf{y} 正交 (orthogonal)，这对应 $E(XY) = 0$ 。

相关性系数的变化

线性相关性系数受到具体样本数据选取的影响。如图 23 所示，对于鸢尾花所有 150 个样本点，花萼长度、花萼宽度的线性相关系数小于 0。但是，分别计算三个不同标签的数据的花萼长度、花萼宽度的线性相关系数，发现这三个值都显著大于 0。

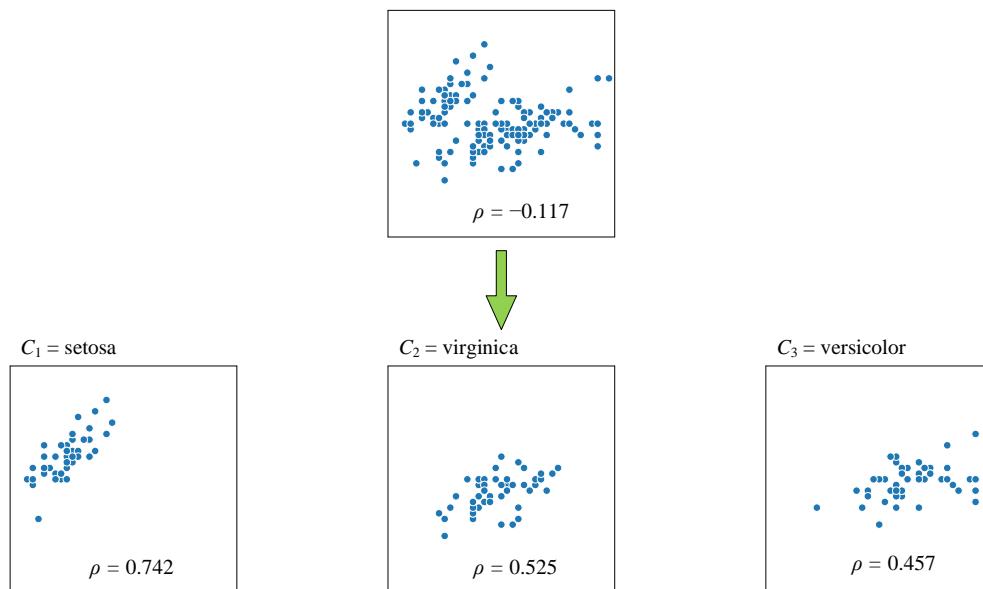


图 23. 鸢尾花不同分类的线性相关性系数

大家将会在《数据有道》一册中看到，如图 24 所示，时间序列数据的相关性系数还会随时间窗口变化。图 24 中，大家看到相关性系数出现陡然上升或下降（高亮）的情况，这可能都是由是几个样本点带来的影响，值得深入研究。

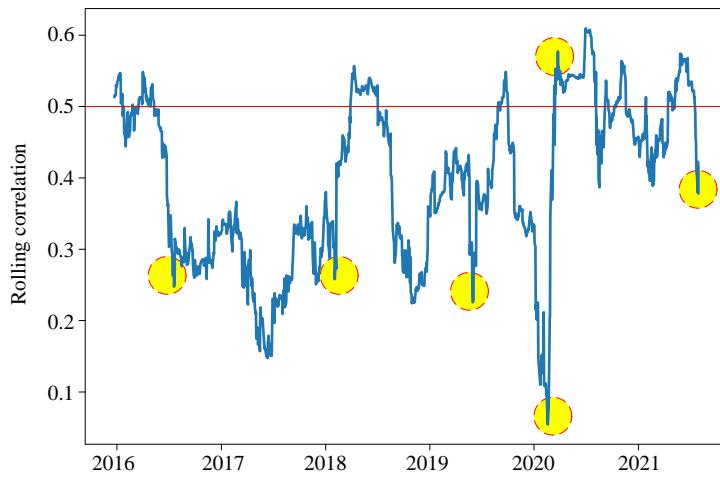


图 24. 移动线性相关性系数

10.6 以鸳尾花数据为例：不考虑分类标签

本节和下一节用二元高斯分布估计鸳尾花萼长度 X_1 、萼片宽度 X_2 的联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 。相信大家还记得我们在本书第 7 章采用 KDE 估计联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 。这两节采用本书和 7 章类似的结构，方便大家比较阅读。

二元高斯分布 → 联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$

假设 (X_1, X_2) 服从二元高斯分布：

$$(X_1, X_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (32)$$

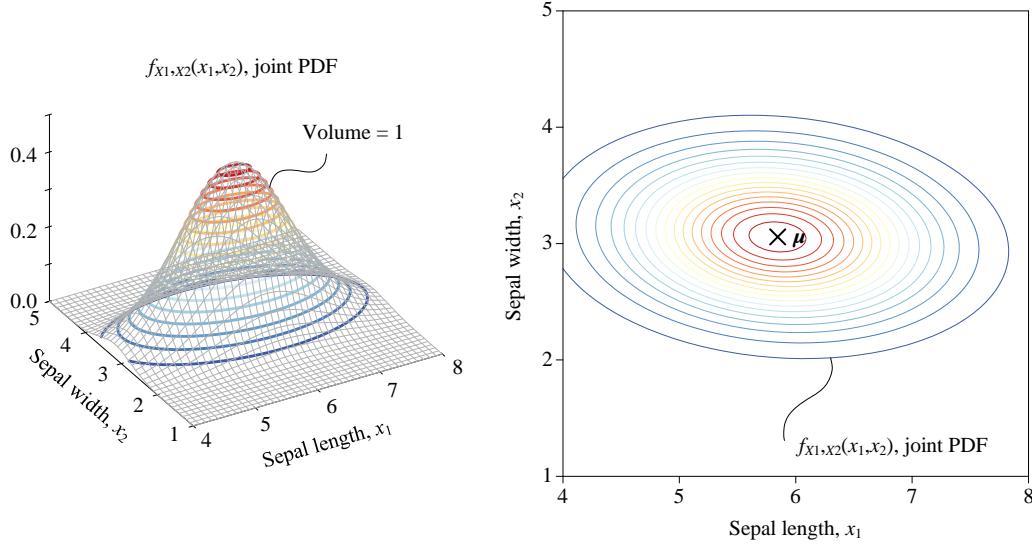
利用鸳尾花 150 个样本数据，我们可以估算得到 (X_1, X_2) 的质心和协方差矩阵分别为：

$$\boldsymbol{\mu} = \begin{bmatrix} 5.843 \\ 3.057 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 0.685 & -0.042 \\ -0.042 & 0.189 \end{bmatrix} \quad (33)$$

(X_1, X_2) 的联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 解析式则为：

$$f_{X_1, X_2}(x_1, x_2) \approx \frac{\exp\left(-\frac{1}{2}(x-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})\right)}{2\pi \times 0.358 \cdot (\sqrt{2\pi})^2 \cdot |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \quad (34)$$

图 25 所示为假设 (X_1, X_2) 服从二元高斯分布时，联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 的三维等高线和平面等高线。

图 25. $f_{x_1, x_2}(x_1, x_2)$ 联合概率密度三维等高线和平面等高线，不考虑分类

举个例子，花萼长度 (X_1) 为 6.5、花萼宽度 (X_2) 为 2.0 时，利用 (34) 估计得到联合概率密度值为：

$$f_{X_1, X_2}(x_1 = 6.5, x_2 = 2.0) \approx 0.0205 \quad (35)$$

注意，这个数值是概率密度，不是概率。但是这个值某种程度上也代表可能性。

马氏距离椭圆的性质

《矩阵力量》第 20 章介绍过**马氏距离** (Mahalanobis distance 或 Mahal distance)，具体定义为：

$$d = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \quad (36)$$

图 26 所示为基于鸢尾花花萼长度、花萼宽度样本数据的马氏距离椭圆。图中，黑色旋转椭圆分别代表马氏距离为 1、2、3、4。图中，还有一个 $\mu_1 \pm \sigma_1$ 和 $\mu_2 \pm \sigma_2$ 构成的矩形。根据本章前文所学，我们知道马氏距离为 1 椭圆和矩形相切于四个点。

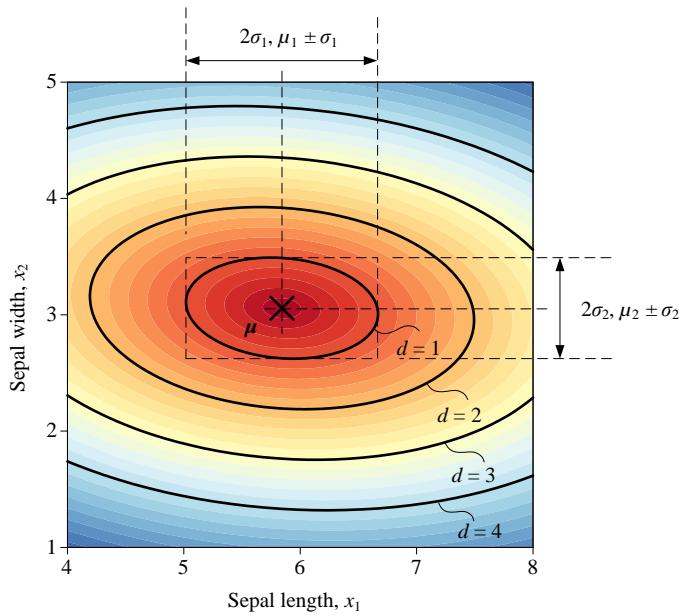


图 26. 马氏距离的椭圆，鸢尾花萼长度、花萼宽度样本数据

还有一个需要大家注意的矩形。如图 27 所示，这个矩形和马氏距离为 1 椭圆同样相切，但是它的长边平行于椭圆的长轴。请大家自行计算椭圆长轴倾斜角。

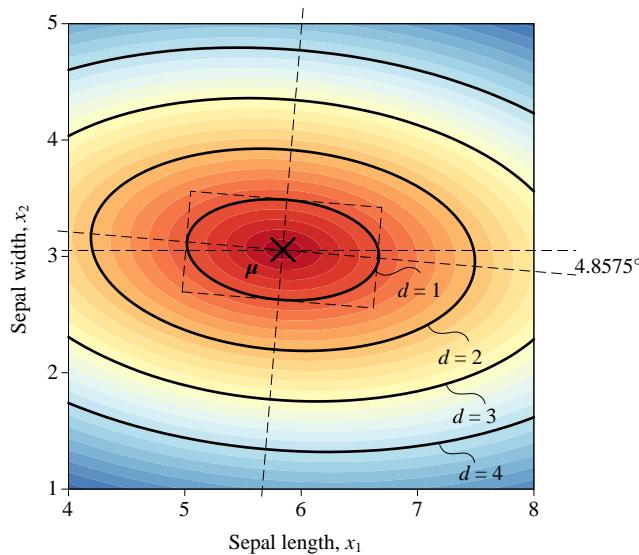


图 27. 马氏距离的椭圆的长轴、短轴，以及对应矩形

我们已经知道马氏距离和概率密度之间的关系为：

$$f_{X_1, X_2}(x_1, x_2) = \frac{\exp\left(-\frac{1}{2}d^2\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \quad (37)$$

对于(34)，当 $d=1$ 时：

$$f_{X_1, X_2}(x_1, x_2)|_{d=1} = \frac{\exp\left(-\frac{1}{2} \times 1^2\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \approx 0.2693 \quad (38)$$

当 $d=2$ 时：

$$f_{X_1, X_2}(x_1, x_2)|_{d=2} = \frac{\exp\left(-\frac{1}{2} \times 2^2\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \approx 0.0601 \quad (39)$$

如图 28 所示，利用二重积分，我们可以计算两幅子图中阴影区域对应的概率：

$$\iint_D f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 \quad (40)$$

从概率统计角度来看，阴影区域有什么意义？这个问题的答案留到本书第 23 章回答。

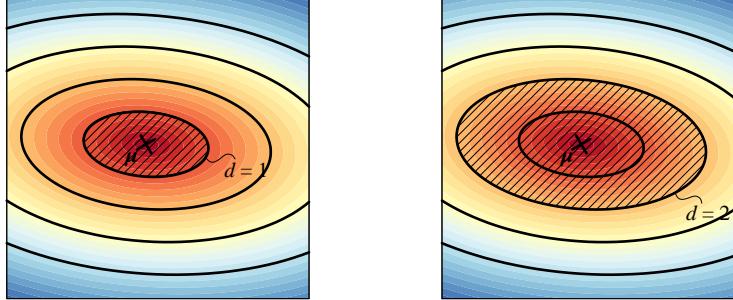
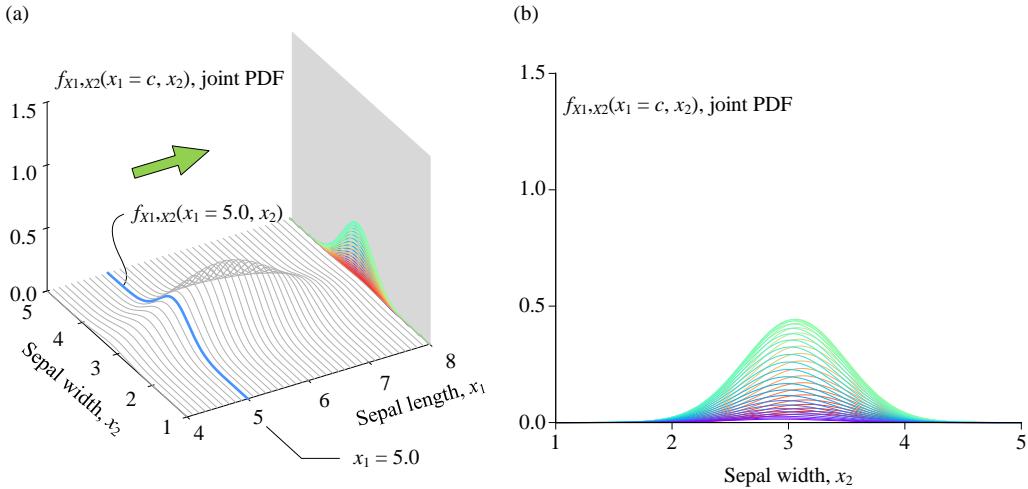
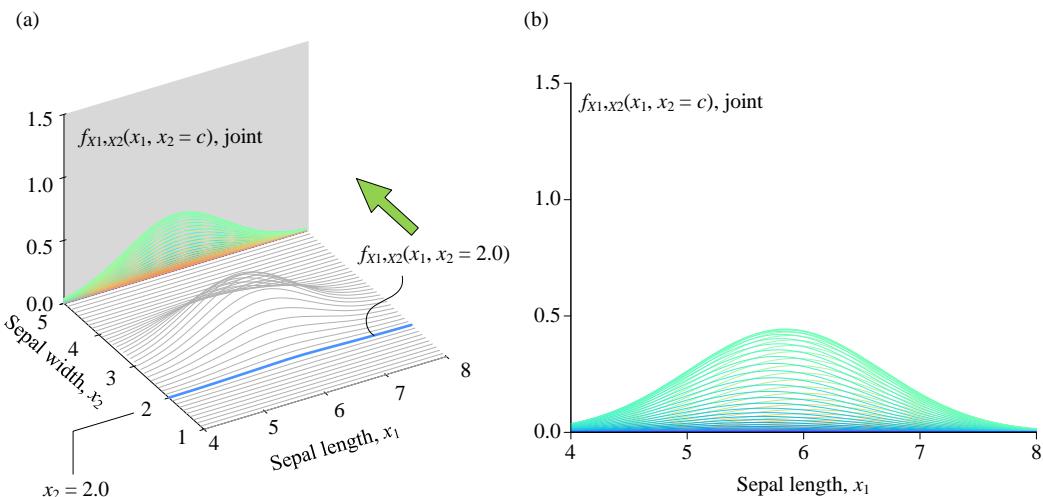


图 28. 求阴影区域对应的概率

联合概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 的剖面线

$f_{X_1, X_2}(x_1, x_2)$ 本质上是个二元函数，因此我们还可以使用“剖面线”分析二元函数。

当固定 x_1 取值时， $f_{X_1, X_2}(x_1 = c, x_2)$ 代表一条曲线。将一系列类似曲线投影到竖直平面得到图 29 (b)。观察图 29 (b)，我们容易发现这些曲线都类似一元高斯分布。图 30 所示为固定 x_2 时，概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 随 x_1 变化。

图 29. 固定 x_1 时，概率密度函数 $f_{X1,X2}(x_1, x_2)$ 随 x_2 变化图 30. 固定 x_2 时，概率密度函数 $f_{X1,X2}(x_1, x_2)$ 随 x_1 变化

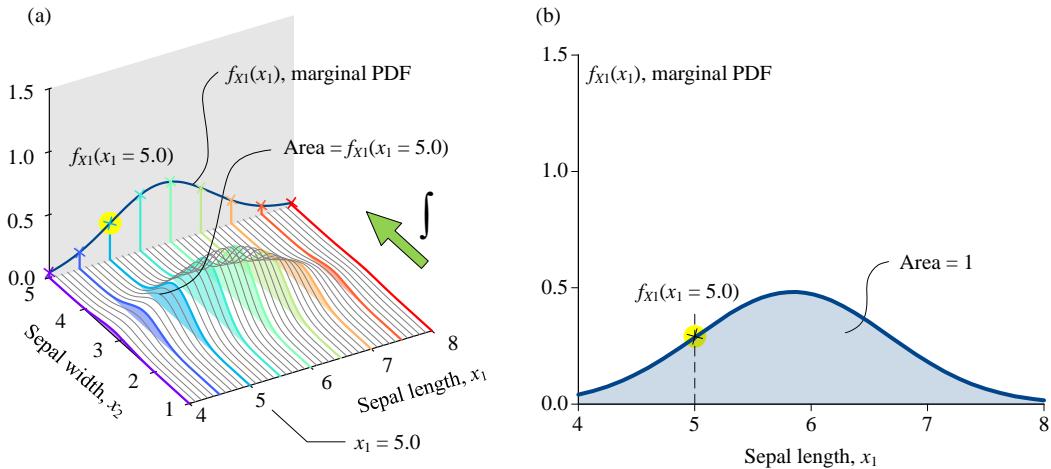
花萼长度边缘概率密度函数 $f_{X1}(x_1)$ ：偏积分

图 31 所示为求解花萼长度边缘概率 $f_{X1}(x_1)$ 的过程：

$$\underbrace{f_{X1}(x_1)}_{\text{Marginal}} = \int_{-\infty}^{+\infty} \underbrace{f_{X1,X2}(x_1, x_2)}_{\text{Joint}} dx_2 \quad (41)$$

图 31 中彩色阴影面积对应边缘概率，即 $f_{X1}(x_1)$ 曲线高度。 $f_{X1}(x_1)$ 本身也是概率密度，不是概率值。 $f_{X1}(x_1)$ 再积分可以得到概率。

如图 31 (b) 所示， $f_{X1}(x_1)$ 曲线和整个横轴围成图形的面积为 1。通过本章前文学习，我们知道 $f_{X1}(x_1)$ 也是一元高斯分布 PDF。

图 31. 偏积分求解边缘概率 $f_{x1}(x1)$

花萼宽度边缘概率 $f_{x2}(x2)$: 偏求和

图 32 所示为求解花萼宽度边缘概率的过程：

$$f_{x2}(x2) = \int_{-\infty}^{+\infty} f_{X1,X2}(x1, x2) dx_1 \quad (42)$$

如所示， $f_{x2}(x2)$ 为一元高斯分布 PDF。

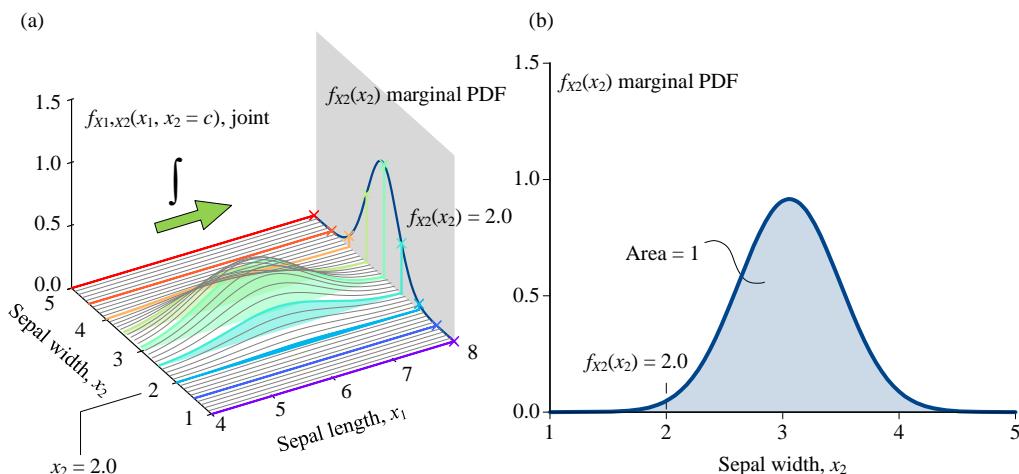
图 32. 偏积分求解边缘概率 $f_{x2}(x2)$

图 33 所示为联合概率和边缘概率之间关系。图中联合概率密度 $f_{X1,X2}(x1, x2)$ 采用二元高斯分布估计得到。图 33 中 $f_{X1,X2}(x1, x2)$ 等高线并没有特别准确捕捉到鸢尾花花萼长度、花萼宽度样本散点分布细节。

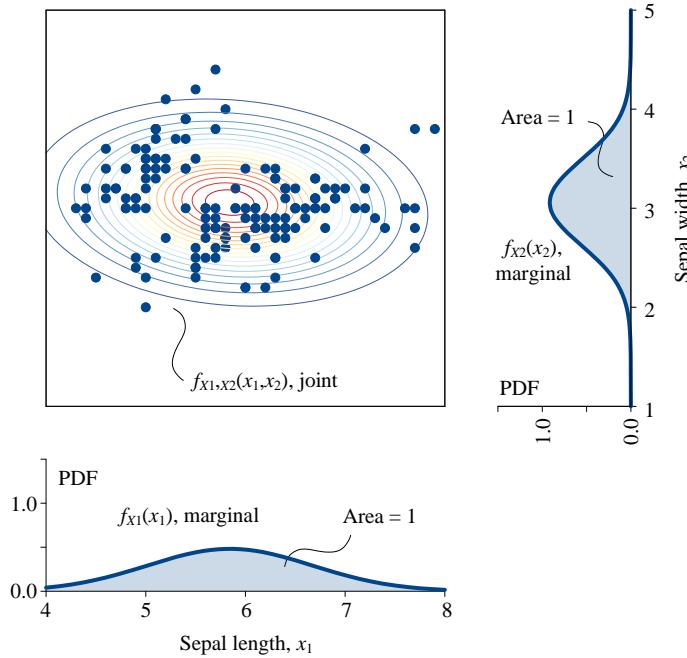


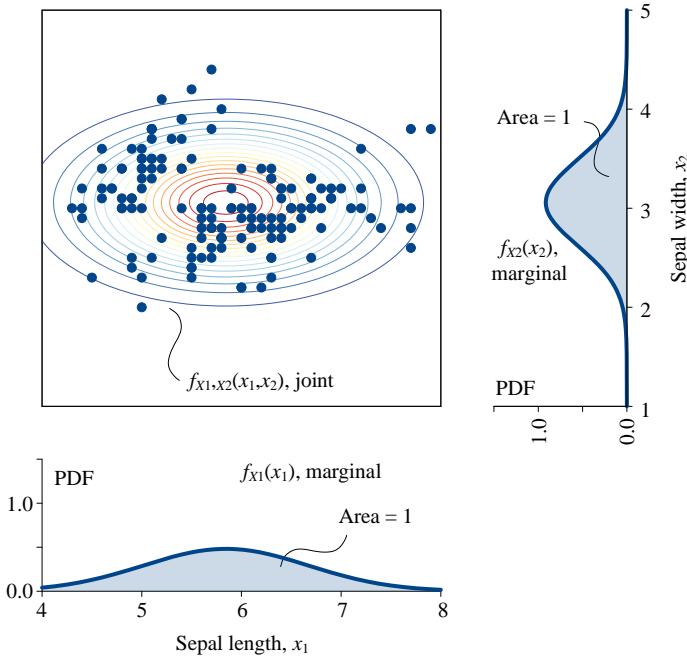
图 33. 联合概率和边缘概率之间关系

假设独立

如果假设 X_1 和 X_2 独立，则联合概率密度 $f_{X_1, X_2}(x_1, x_2)$ 可通过下式计算得到：

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \quad (43)$$

图 34 所示为假设 X_1 和 X_2 独立时 $f_{X_1, X_2}(x_1, x_2)$ 的平面等高线和边缘概率之间关系。椭圆等高线为正椭圆，而非旋转椭圆（图 33）。

图 34. 联合概率，假设 X_1 和 X_2 独立

给定花萼长度，花萼宽度的条件概率密度 $f_{X2|X1}(x_2 | x_1)$

如图 35 所示，利用贝叶斯定理，条件概率密度 $f_{X2|X1}(x_2 | x_1)$ 可以通过下式计算：

$$\underbrace{f_{X2|X1}(x_2 | x_1)}_{\text{Conditional}} = \frac{\overbrace{f_{X1,X2}(x_1, x_2)}^{\text{Joint}}}{\underbrace{f_{X1}(x_1)}_{\text{Marginal}}} \quad (44)$$

分母中的边缘概率 $f_{X1}(x_1) (>0)$ 起到归一化作用。如图 35 (b) 所示，经过归一化的条件概率曲线围成的面积变为 1。

将不同位置的条件概率密度 $f_{X2|X1}(x_2 | x_1)$ 曲线投影到平面得到图 36。我们隐约发现图 36 (b) 中每条曲线看上去都是一元高斯分布。

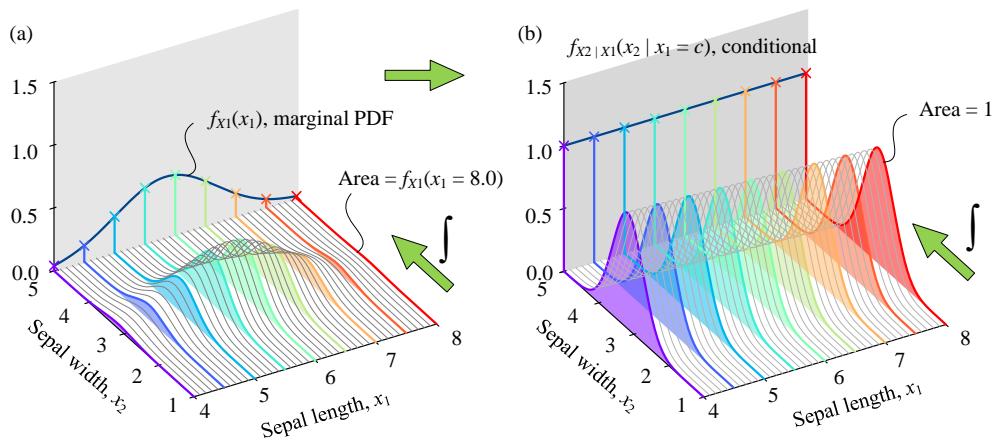
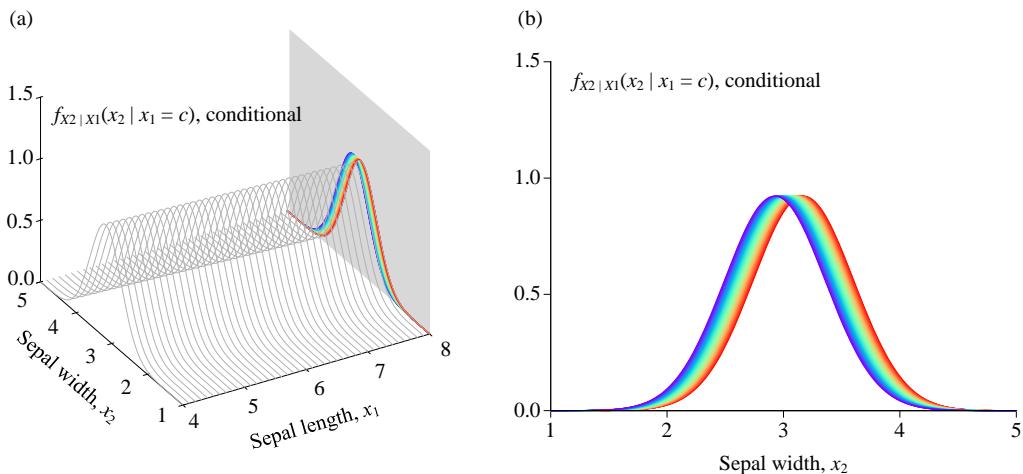
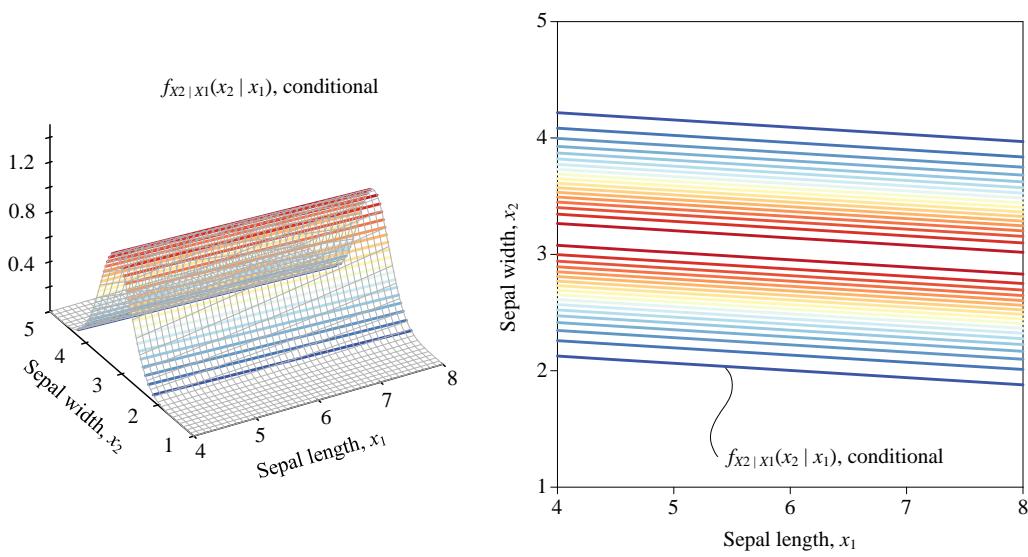
这难道是个巧合？

我们将在本书第 13 章揭晓答案。

$f_{X2|X1}(x_2 | x_1)$ 本身也是一个二元函数。图 37 所示为 $f_{X2|X1}(x_2 | x_1)$ 三维等高线和平面等高线。从平面等高线中，我们看到一系列直线。

这难道也是个巧合？

答案同样在本书第 13 章给出。

图 35. 计算条件概率 $f_{X2|X1}(x_2|x_1)$ 原理图 36. $f_{X2|X1}(x_2|x_1)$ 曲线投影到平面图 37. $f_{X2|X1}(x_2|x_1)$ 条件概率密度三维等高线和平面等高线，不考虑分类

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

给定花萼宽度，花萼长度的条件概率密度函数 $f_{X_1|X_2}(x_1|x_2)$

如图 38 所示，同样利用贝叶斯定理，条件概率密度 $f_{X_1|X_2}(x_1|x_2)$ 可以通过下式计算：

$$\underbrace{f_{X_1|X_2}(x_1|x_2)}_{\text{Conditional}} = \frac{\overbrace{f_{X_1, X_2}(x_1, x_2)}^{\text{Joint}}}{\underbrace{f_{X_2}(x_2)}_{\text{Marginal}}} \quad (45)$$

类似前文，上式中分母中 $f_{X_2}(x_2) (> 0)$ 起到归一化作用。

将不同位置的条件概率密度 $f_{X_1|X_2}(x_1|x_2)$ 曲线投影到平面得到图 39。图 39 (b) 中每条曲线也都类似一元高斯分布曲线。

$f_{X_1|X_2}(x_1|x_2)$ 同样也是一个二元函数，如图 40 所示的 $f_{X_1|X_2}(x_1|x_2)$ 三维等高线和平面等高线。

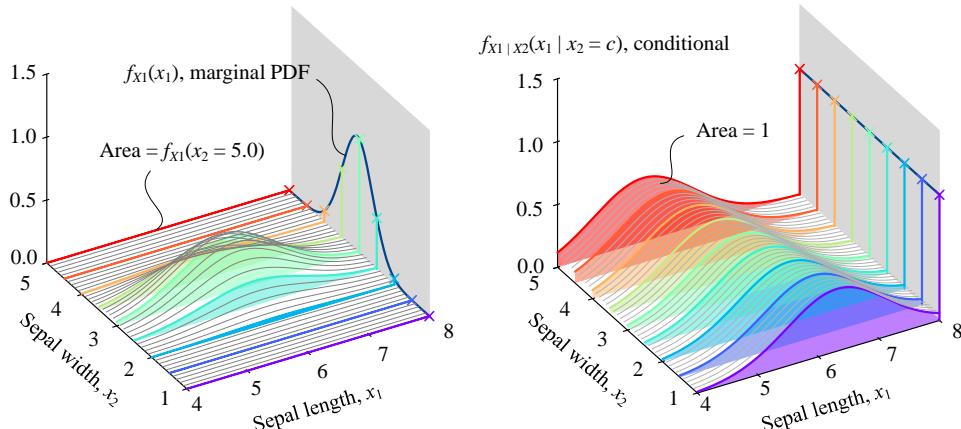


图 38. 计算条件概率 $f_{X_1|X_2}(x_1|x_2)$ 原理

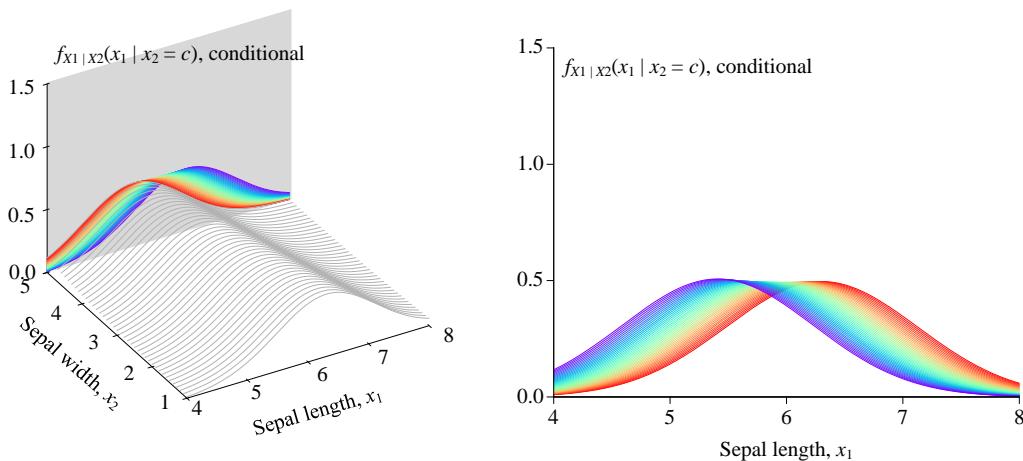
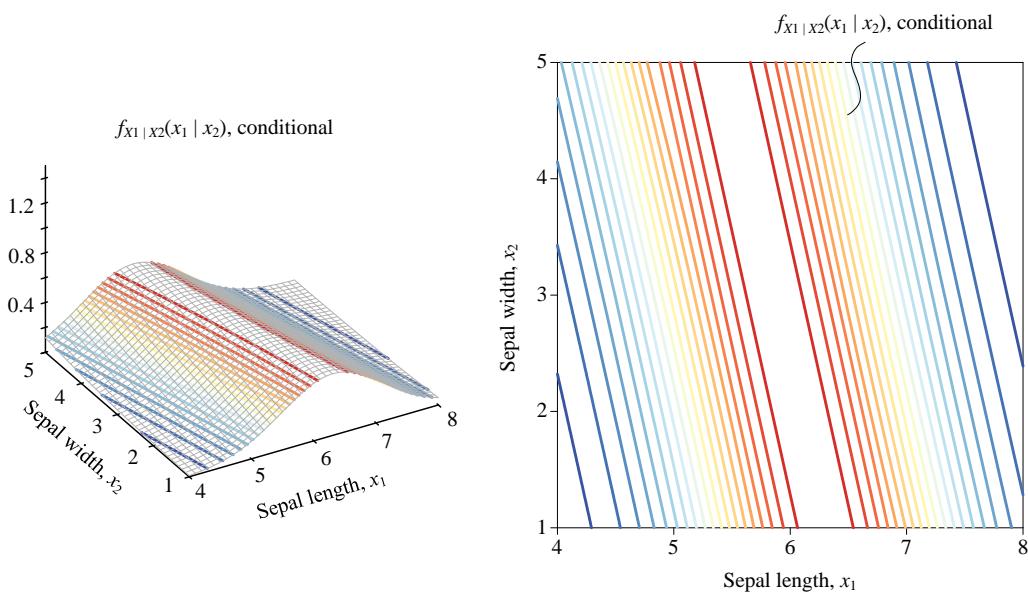


图 39. $f_{X_1|X_2}(x_1|x_2)$ 曲线投影到平面

图 40. $f_{x_1|x_2}(x_1 | x_2)$ 条件概率密度三维等高线和平面等高线，不考虑分类

10.7 以鸢尾花数据为例：考虑分类标签

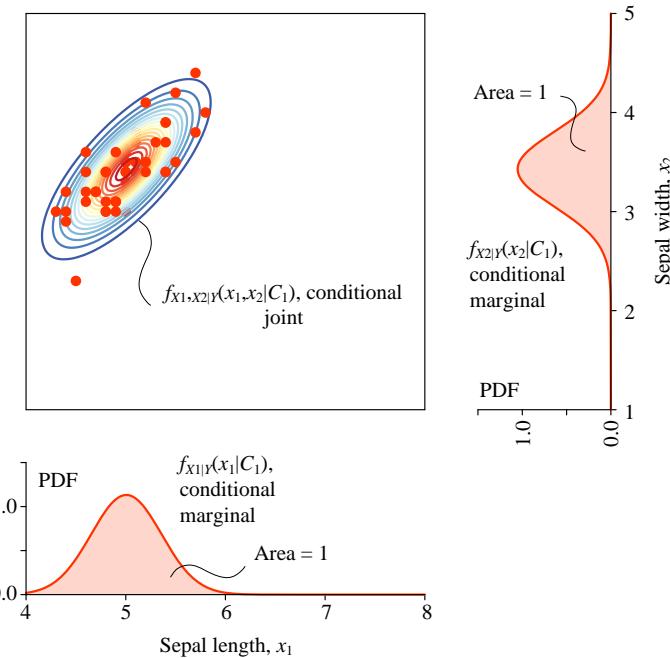
本节讨论考虑鸢尾花分类条件下的条件概率 PDF。

给定分类标签 $Y = C_1$ (setosa)

给定分类标签 $Y = C_1$ (setosa) 条件下，假设鸢尾花萼长度、萼片宽度同样服从二元高斯分布。

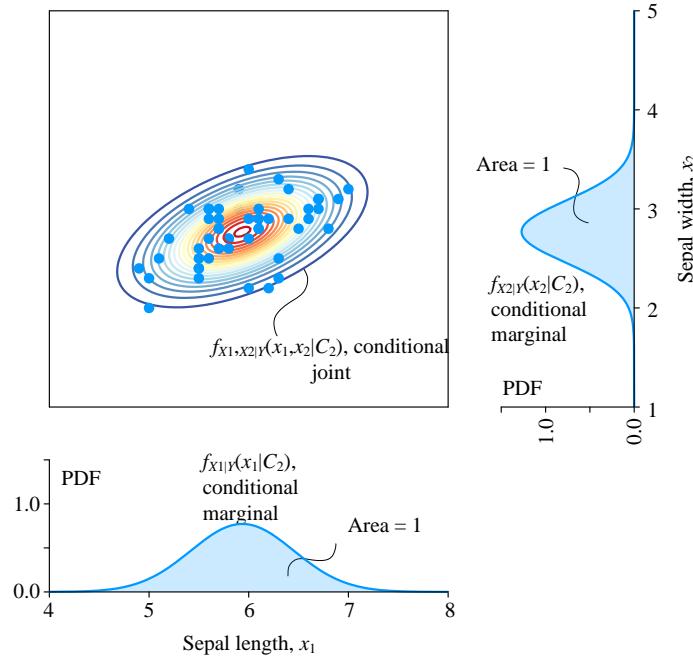
图 41 所示为，给定分类标签 $Y = C_1$ (setosa)，条件概率 $f_{x_1, x_2 | Y}(x_1, x_2 | y = C_1)$ 平面等高线和条件边缘概率密度曲线。 $f_{x_1, x_2 | Y}(x_1, x_2 | y = C_1)$ 曲面和整个水平面围成体积为 1。

图 41 中 $f_{x_1 | Y}(x_1 | y = C_1)$ 、 $f_{x_2 | Y}(x_2 | y = C_1)$ 分别和 x_1 、 x_2 围成的面积也是 1。

图 41. 条件概率 $f_{X_1, X_2|Y}(x_1, x_2|y = C_1)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_1$ (setosa)

给定分类标签 $Y = C_2$ (versicolor)

图 42 所示为，给定分类标签 $Y = C_2$ (versicolor)，条件概率 $f_{X_1, X_2|Y}(x_1, x_2|y = C_2)$ 平面等高线和条件边缘概率密度曲线。

图 42. 条件概率 $f_{X_1, X_2|Y}(x_1, x_2|y = C_2)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_2$ (versicolor)

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

给定分类标签 $Y = C_3$ (virginica)

图 43 所示为，给定分类标签 $Y = C_3$ (virginica)，条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_3)$ 平面等高线和条件边缘概率密度曲线。

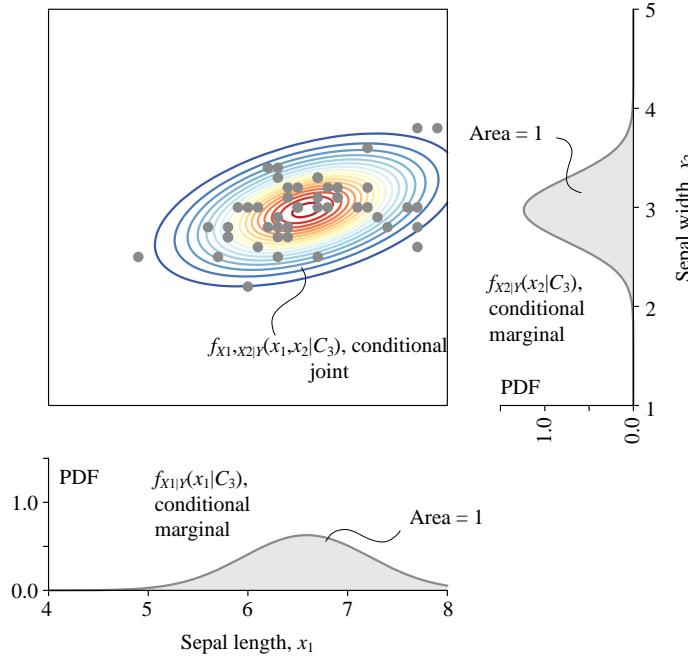


图 43. 条件概率 $p_{X_1, X_2|Y}(x_1, x_2 | y = C_3)$ 平面等高线和条件边缘概率密度曲线，给定分类标签 $Y = C_3$ (virginica)

全概率

如图 44 所示，利用全概率定理，三个条件概率等高线叠加可以得到联合概率密度，即：

$$\begin{aligned} f_{X_1, X_2}(x_1, x_2) &= f_{X_1, X_2|Y}(x_1, x_2 | y = C_1) p_Y(C_1) + \\ &\quad f_{X_1, X_2|Y}(x_1, x_2 | y = C_2) p_Y(C_2) + \\ &\quad f_{X_1, X_2|Y}(x_1, x_2 | y = C_3) p_Y(C_3) \end{aligned} \quad (46)$$

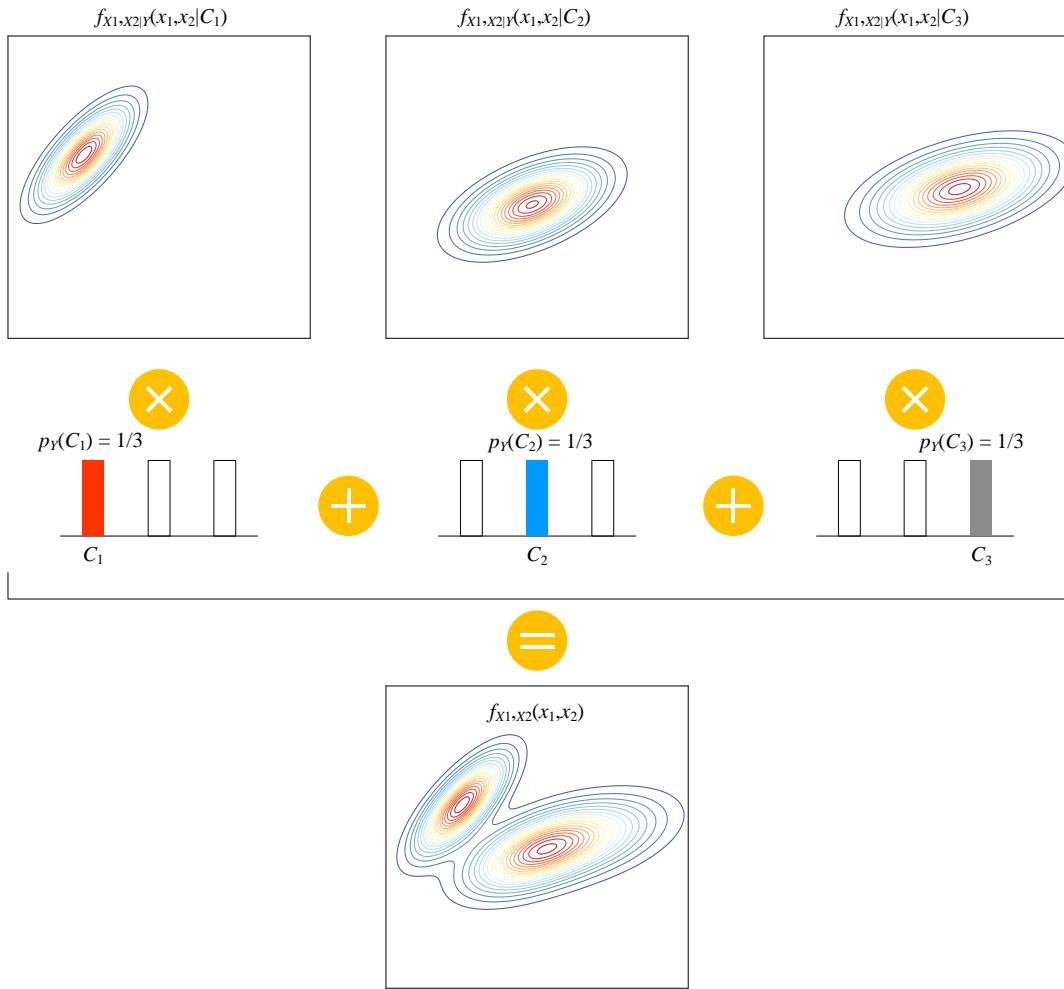


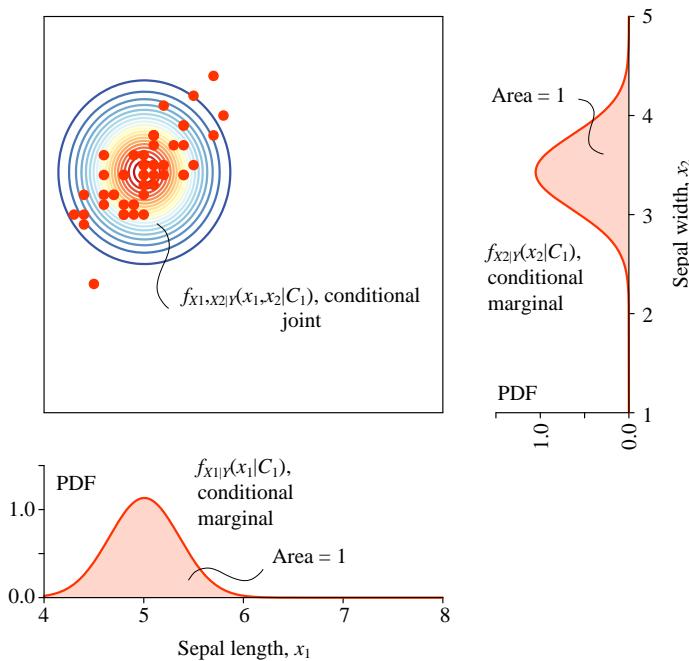
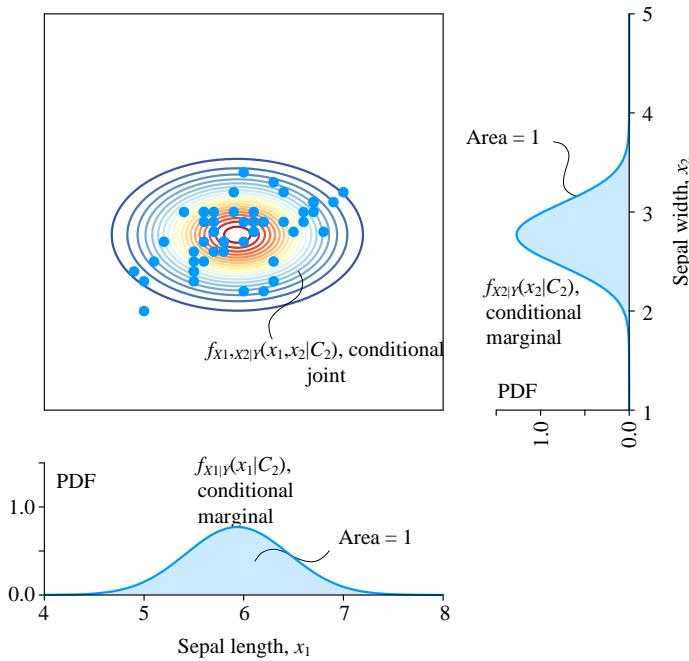
图 44. 估算联合概率密度，假设条件概率服从二元高斯分布

假设条件独立

如图 45 所示，如果假设条件独立， $f_{X1,X2|Y}(x_1, x_2 | y = C_1)$ 可以通过下式计算得到：

$$\underbrace{f_{X1,X2|Y}(x_1, x_2 | y = C_1)}_{\text{Conditional joint}} = \underbrace{f_{X1|Y}(x_1 | y = C_1)}_{\text{Conditional marginal}} \cdot \underbrace{f_{X2|Y}(x_2 | y = C_1)}_{\text{Conditional marginal}} \quad (47)$$

同理我们可以计算得到 $f_{X1,X2|Y}(x_1, x_2 | y = C_2)$ 、 $f_{X1,X2|Y}(x_1, x_2 | y = C_3)$ ，具体如图 46、图 47 所示。

图 45. 给定 $Y = C_1$, X_1 和 X_2 条件独立, 估算条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_1)$ 图 46. 给定 $Y = C_2$, X_1 和 X_2 条件独立, 估算条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_2)$

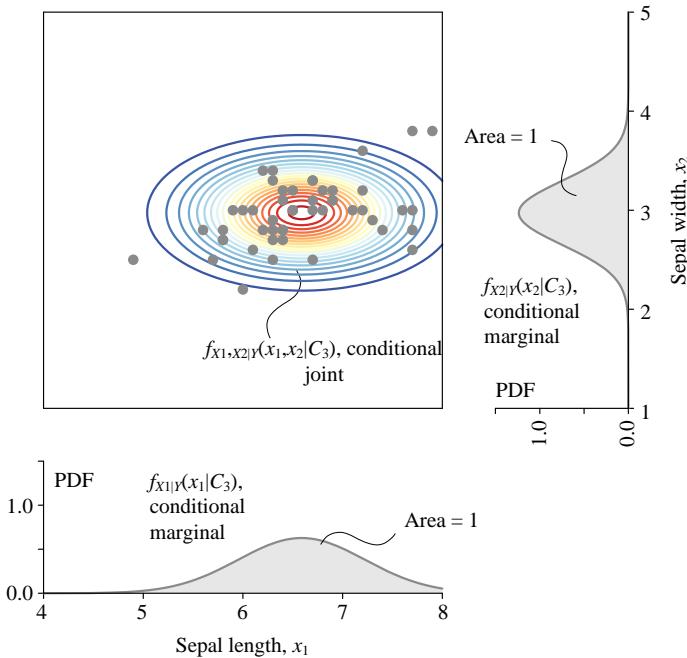
本 PDF 文件为作者草稿, 发布目的为方便读者在移动终端学习, 终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有, 请勿商用, 引用请注明出处。

代码及 PDF 文件下载: <https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger: <https://space.bilibili.com/513194466>

欢迎大家批评指教, 本书专属邮箱: jiang.visualize.ml@gmail.com

图 47. 给定 $Y = C_3$, X_1 和 X_2 条件独立, 估算条件概率 $f_{X1,X2|Y}(x_1, x_2 | y = C_3)$

估计联合概率

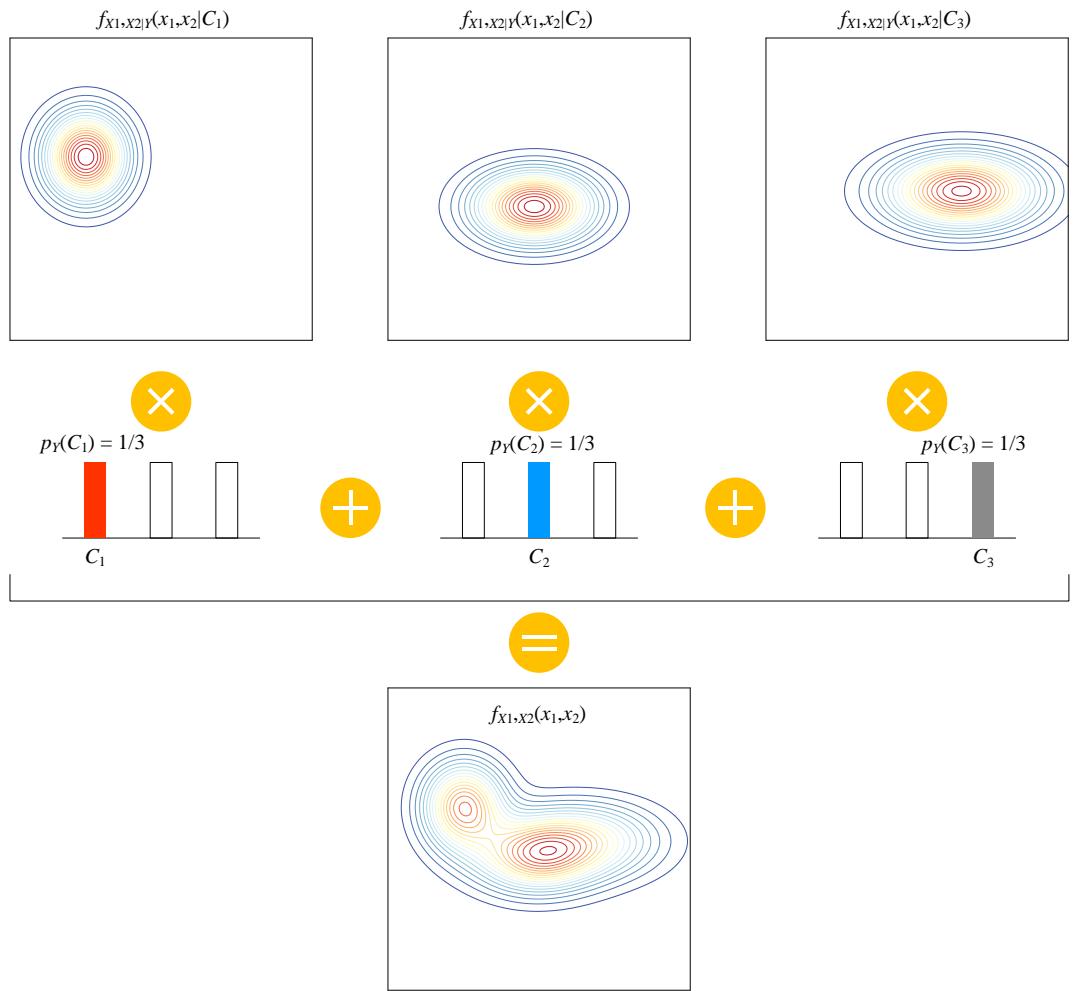
如图 48 所示, 在假设条件独立情况下, 利用全概率定理估算 $f_{X1,X2}(x_1, x_2)$:

$$\begin{aligned} f_{X1,X2}(x_1, x_2) &= f_{X1|Y}(x_1 | y = C_1) f_{X2|Y}(x_2 | y = C_1) p_Y(C_1) + \\ &\quad f_{X1|Y}(x_1 | y = C_2) f_{X2|Y}(x_2 | y = C_2) p_Y(C_2) + \\ &\quad f_{X1|Y}(x_1 | y = C_3) f_{X2|Y}(x_2 | y = C_3) p_Y(C_3) \end{aligned} \quad (48)$$

图 44 和图 48 涉及的这些技术细节对于理解贝叶斯分类器原理有重要意义。



本书第 19、20 章将从贝叶斯定理视角简单介绍分类原理, 《机器学习》一册将专门讲解朴素贝叶斯分类器。

图 48. 利用全概率定理，估算 $f_{X1,X2}(x_1, x_2)$ ，假设条件独立

二元高斯分布的概率密度函数的等高线呈现出椭圆形状，这一点极其重要。这个椭圆将把协方差矩阵、特征值分解、Cholesky 分解、条件概率、马氏距离、线性回归、主成分分析、高斯混合模型、高斯过程等一系列概念紧密联系起来。

11

Multivariate Gaussian Distribution

多元高斯分布

几何、代数、概率统计的完美结合



在我看来，数学科学是一个不可分割的有机体，其生命力取决于各部分的联系。

Mathematical science is in my opinion an indivisible whole, an organism whose vitality is conditioned upon the connection of its parts.

—— 大卫·希尔伯特 (David Hilbert) | 德国数学家 | 1862 ~ 1943



- ▶ `numpy.cov()` 计算协方差矩阵
- ▶ `numpy.diag()` 如果 A 为方阵, `numpy.diag(A)` 函数提取对角线元素, 以向量形式输入结果; 如果 a 为向量, `numpy.diag(a)` 函数将向量展开成方阵, 方阵对角线元素为 a 向量元素
- ▶ `numpy.linalg.eig()` 特征值分解
- ▶ `numpy.linalg.inv()` 计算逆矩阵
- ▶ `numpy.linalg.norm()` 计算范数
- ▶ `numpy.linalg.svd()` 奇异值分解
- ▶ `scipy.spatial.distance.euclidean()` 计算欧氏距离
- ▶ `scipy.spatial.distance.mahalanobis()` 计算马氏距离
- ▶ `seaborn.heatmap()` 绘制热图
- ▶ `seaborn.kdeplot()` 绘制 KDE 核概率密度估计曲线
- ▶ `seaborn.pairplot()` 绘制成对分析图
- ▶ `sklearn.decomposition.PCA()` 主成分分析函数



11.1 矩阵角度：一元、二元、三元到多元

一元

本书第 9 章讲解了一元高斯分布的 PDF 解析式，具体如下：

$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (1)$$

图 1 (a) 所示为一元高斯分布 PDF 的图像。

二元

第 10 章中，我们看到二元高斯分布的 PDF 解析式：

$$f_{x,y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{x,y}^2}} \times \exp\left(-\frac{1}{2} \underbrace{\frac{1}{(1-\rho_{x,y}^2)} \left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho_{x,y} \left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 \right]}_{\text{Ellipse}}\right) \quad (2)$$

图 1 (b) 所示为二元高斯分布 PDF 的图像。

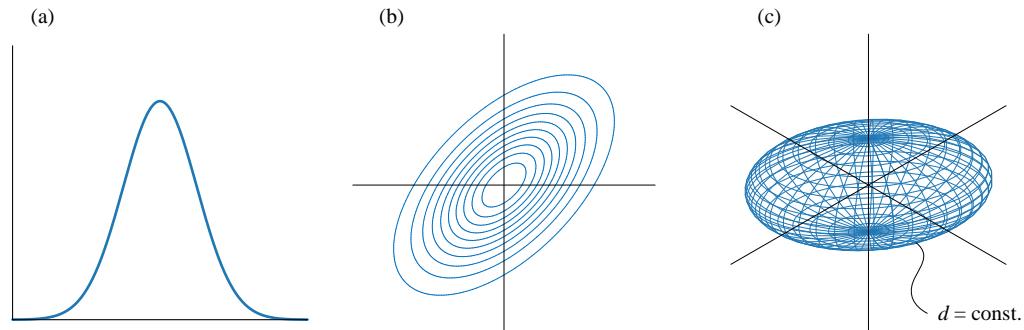


图 1. 一元、二元、三元高斯分布的几何形态

三元

(2) 已经很复杂，我们再看看三元高斯分布 PDF 解析式。在 $\sigma_1 = \sigma_2 = \sigma_3 = 1$, $\mu_1 = \mu_2 = \mu_3 = 0$ 条件下，三元高斯分布 PDF 解析式如下：

$$f_{x_1,x_2,x_3}(x_1, x_2, x_3) = \frac{\exp\left(-\frac{1}{2}d^2\right)}{(2\pi)^{\frac{3}{2}} \sqrt{1+2\rho_{1,2}\rho_{1,3}\rho_{2,3}-(\rho_{1,2}^2+\rho_{1,3}^2+\rho_{2,3}^2)}} \quad (3)$$

其中

$$d^2 = \frac{x_1^2(\rho_{2,3}^2 - 1) + x_2^2(\rho_{1,3}^2 - 1) + x_3^2(\rho_{1,2}^2 - 1) + 2[x_1x_2(\rho_{1,2} - \rho_{1,3}\rho_{2,3}) + x_1x_3(\rho_{1,3} - \rho_{1,2}\rho_{2,3}) + x_2x_3(\rho_{2,3} - \rho_{1,3}\rho_{1,2})]}{(\rho_{1,2}^2 + \rho_{1,3}^2 + \rho_{2,3}^2 - 2\rho_{1,2}\rho_{1,3}\rho_{2,3} - 1)} \quad (4)$$

当 d 为确定值时，上式代表一个椭球 (ellipsoid)，如图 1 (c) 所示。也就是说三元高斯分布 PDF 的几何图形是嵌套的椭球。

相信大家已经看到了三元高斯分布 PDF 解析式的复杂程度。更不用说，(3) 的解析式是在 $\sigma_1 = \sigma_2 = \sigma_3 = 1, \mu_1 = \mu_2 = \mu_3 = 0$ 这个极特殊条件下获得的。

到了四元、五元、更高元高斯分布 PDF 解析式时，代数展开式已经完全不够用了。因此，对于多元高斯分布，我们需要矩阵算式。

多元

“鸢尾花书”读者应该已经很熟悉多元正态分布 PDF，具体如下：

$$f_\chi(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \quad (5)$$

其中， χ 、 \mathbf{x} 、 $\boldsymbol{\mu}$ 均为列向量：

$$\chi = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} \quad (6)$$

向量 $\boldsymbol{\mu}$ 常常被称作质心 (centroid)， D 为高斯分布的特征数，比如二元高斯分布 $D = 2$ 。

协方差矩阵 $\boldsymbol{\Sigma}$ 为：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D,1} & \sigma_{D,2} & \cdots & \sigma_{D,D} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,D}\sigma_1\sigma_D \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,D}\sigma_2\sigma_D \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D}\sigma_1\sigma_D & \rho_{2,D}\sigma_2\sigma_D & \cdots & \sigma_D^2 \end{bmatrix} \quad (7)$$

⚠ 特别需要大家注意的是，如果 (5) 成立，协方差矩阵 $\boldsymbol{\Sigma}$ 必须为正定矩阵。如果为 $\boldsymbol{\Sigma}$ 半正定， $\boldsymbol{\Sigma}$ 的行列式值为 0，而 (5) 分母不能为 0。 $\boldsymbol{\Sigma}$ 半正定说明 χ 存在线性相关。

一组随机变量构成的列向量 χ 服从如 (5) 多元高斯分布，记做：

$$\chi = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix}, \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D,1} & \sigma_{D,2} & \cdots & \sigma_{D,D} \end{bmatrix}\right) \quad (8)$$

或更简便地记做：

$$\chi \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (9)$$

再次强调，这个语境下， χ 为随机变量构成的列向量，每一行代表一个随机变量；而 X 代表数据矩阵，每一列对应一个随机变量所有样本。

多元 → 一元

$D = 1$ 时，质心为：

$$\boldsymbol{\mu} = [\mu] \quad (10)$$

协方差矩阵为：

$$\boldsymbol{\Sigma} = [\sigma^2] \quad (11)$$

(5) 分子中的**二次型** (quadratic form) 可以展开为：

$$(x - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (x - \boldsymbol{\mu}) = (x - \boldsymbol{\mu}) \boldsymbol{\sigma}^{-2} (x - \boldsymbol{\mu}) = \left(\frac{x - \boldsymbol{\mu}}{\sigma} \right)^2 \quad (12)$$

我们看到的是 Z 分数的平方。这和 (1) 解析式完全一致。

多元 → 二元

再以二元 ($D = 2$) 高斯分布为例，它的质心：

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad (13)$$

二元高斯分布的协方差矩阵 $\boldsymbol{\Sigma}$ 具体为：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (14)$$

协方差矩阵的行列式值 $|\boldsymbol{\Sigma}|$ ：

$$|\boldsymbol{\Sigma}| = \sigma_1^2 \sigma_2^2 - \rho_{1,2}^2 \sigma_1^2 \sigma_2^2 = \sigma_1^2 \sigma_2^2 (1 - \rho_{1,2}^2) \quad (15)$$

再次强调，如果相关性系数为 ± 1 ，行列式值 $|\Sigma|$ 为0。相关性系数取值范围为 $(-1, 1)$ 时，协方差矩阵的逆 Σ^{-1} 为：

$$\Sigma^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2 (1 - \rho_{1,2}^2)} \begin{bmatrix} \sigma_2^2 & -\rho_{1,2} \sigma_1 \sigma_2 \\ -\rho_{1,2} \sigma_1 \sigma_2 & \sigma_1^2 \end{bmatrix} = \frac{1}{1 - \rho_{1,2}^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho_{1,2}}{\sigma_1 \sigma_2} \\ \frac{-\rho_{1,2}}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \quad (16)$$

对于二元高斯分布，(5) 分子中的二次型展开：

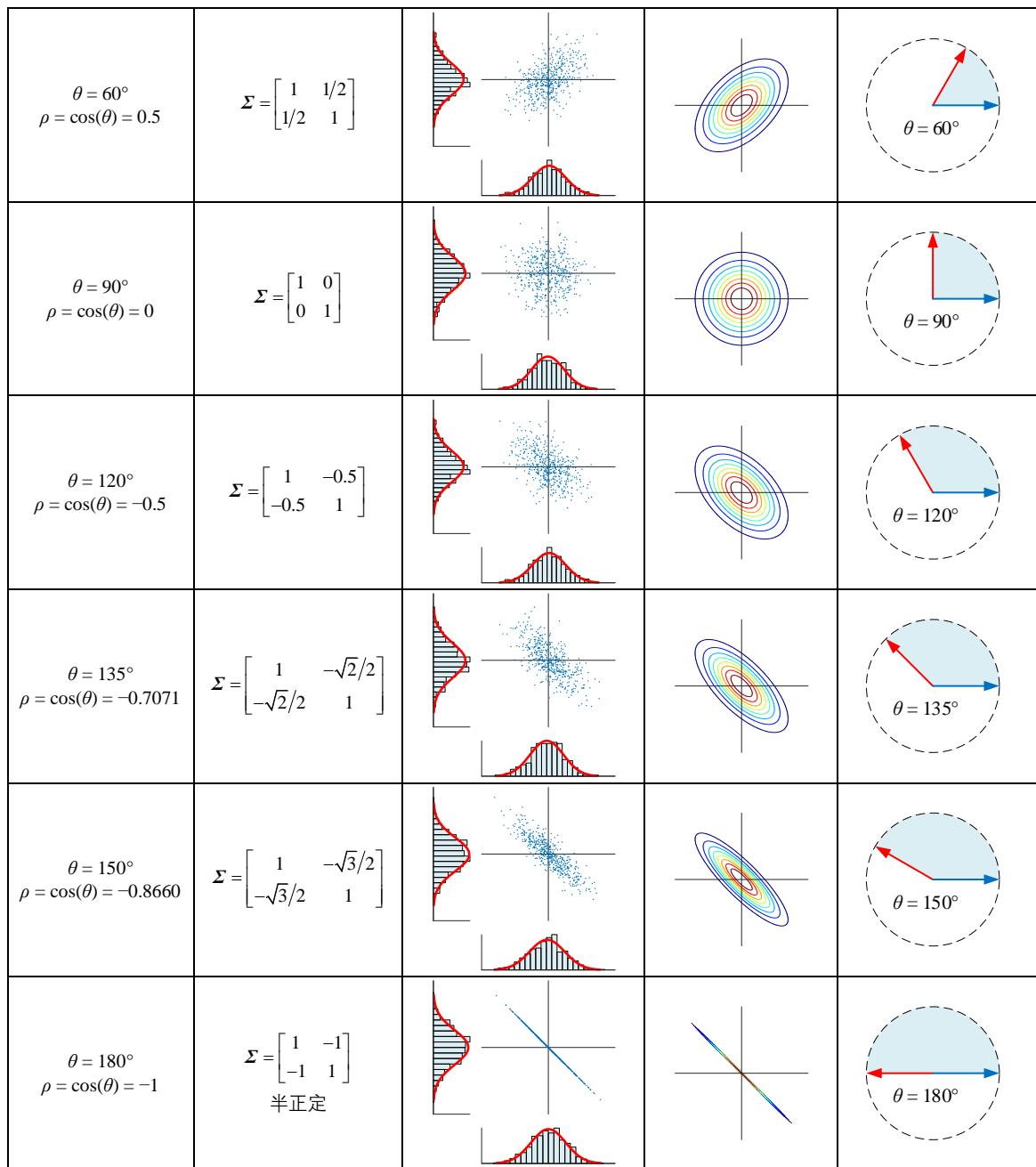
$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) &= [x_1 - \mu_1 \quad x_2 - \mu_2] \frac{1}{1 - \rho_{1,2}^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho_{1,2}}{\sigma_1 \sigma_2} \\ \frac{-\rho_{1,2}}{\sigma_1 \sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\ &= \frac{1}{1 - \rho_{1,2}^2} \left[\left(\frac{x_1 - \mu_1}{\sigma_1} \right)^2 - 2\rho_{1,2} \left(\frac{x_1 - \mu_1}{\sigma_1} \right) \left(\frac{x_2 - \mu_2}{\sigma_2} \right) + \left(\frac{x_2 - \mu_2}{\sigma_2} \right)^2 \right] \end{aligned} \quad (17)$$

分别将(17)和(15)代入(5)可以得到二元高斯分布 PDF 解析式。

表 1 所示为不同线性相关系数的可视化方案。

表 1. 不同相关性系数的可视化方案

相关性系数	协方差矩阵	散点图	PDF 等高线	向量
$\theta = 0^\circ$ $\rho = \cos(\theta) = 1$	$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ 半正定			
$\theta = 30^\circ$ $\rho = \cos(\theta) = 0.8660$	$\Sigma = \begin{bmatrix} 1 & \sqrt{3}/2 \\ \sqrt{3}/2 & 1 \end{bmatrix}$			
$\theta = 45^\circ$ $\rho = \cos(\theta) = 0.7071$	$\Sigma = \begin{bmatrix} 1 & \sqrt{2}/2 \\ \sqrt{2}/2 & 1 \end{bmatrix}$			



随机变量独立

特别地，如果 (X_1, X_2) 服从二元高斯分布，并且随机变量 X_1 和 X_2 独立，这样 (X_1, X_2) 的协方差矩阵为：

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad (18)$$

注意，这个协方差矩阵为对角阵。

根据上一章所学，我们知道 X_1 和 X_2 各自的边缘概率密度函数分别为：

$$\begin{aligned} f_{X_1}(x_1) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right) \\ f_{X_2}(x_2) &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right) \end{aligned} \quad (19)$$

如果 (X_1, X_2) 服从二元高斯函数，且 X_1 和 X_2 独立， (X_1, X_2) 概率密度函数可以写成两个边缘概率密度函数的乘积：

$$\begin{aligned} \underbrace{f_{X_1, X_2}(x_1, x_2)}_{\text{Joint}} &= \frac{1}{2\pi\sigma_1\sigma_2} \times \exp\left(-\frac{1}{2}\left(\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2 + \left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right)\right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{x_1 - \mu_1}{\sigma_1}\right)^2\right)}_{\text{Marginal}, f_{X_1}(x_1)} \times \underbrace{\frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{x_2 - \mu_2}{\sigma_2}\right)^2\right)}_{\text{Marginal}, f_{X_2}(x_2)} \end{aligned} \quad (20)$$

这种情况，二元高斯分布 PDF 等高线为正椭圆。

11.2 高斯分布：椭圆、椭球、超椭球

椭圆分布

上一章提过高斯分布是**椭圆分布** (elliptical distribution) 的一种特殊形式。而椭圆分布的 PDF 一般形式为：

$$f(\mathbf{x}) = k \cdot g \left[\underbrace{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}_{\text{Ellipse}} \right] \quad (21)$$

本书第 7 章介绍的学生 t -分布、逻辑分布、拉普拉斯分布也都是椭圆分布家族成员。

二元高斯分布：椭圆结构

回顾上一章介绍的二元高斯分布的椭圆结构。如图 2 所示，椭圆中心对应质心 $\boldsymbol{\mu}$ ，椭圆和 $\pm\sigma$ 标准差构成的长方形相切，四个切点分别为 A 、 B 、 C 和 D ，对角切点两两相连得到两条直线 AC 、 BD 。

AC 相当于在给定 X_2 条件下 X_1 的条件概率期望值； BD 相当于在给定 X_1 条件下 X_2 的条件概率期望值，这是本书第 12 章要讨论的话题。

在椭圆的学习中，我们很关注椭圆的长轴、短轴，对应图 2 中两条红线 EG 、 FH 。 EG 通过椭圆圆心 O 最长的线段，为椭圆长轴； FH 通过椭圆中心 O 最短的线段，为椭圆短轴。获得长轴、短轴的长度、角度需要用到特征值分解，这是本章后续要讨论的内容。



而长轴就是主成分分析的第一主元方向，这是本书第 14、25 章要讨论的话题。

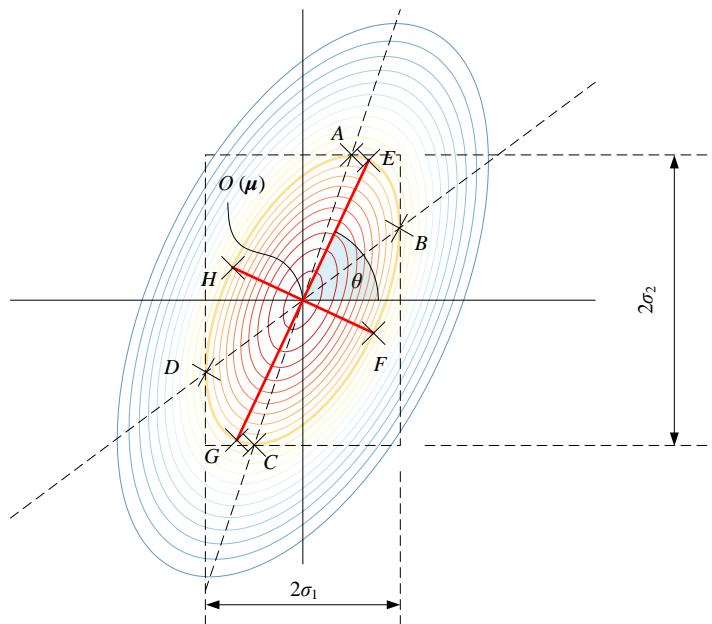


图 2. 椭圆和 $\pm\sigma$ 标准差长方形的关系



Bk5_Ch11_01.py 绘制图 2。

三元高斯分布

前文提过，三元高斯分布 PDF 的几何图形是一层层“嵌套”的椭球。为了看见三元高斯分布 PDF 的椭球，我们采用“切片”这种可视化方案。



《可视之美》一册介绍过这种可视化方案。

图 3 可视化三元高斯分布的 PDF，这个高斯分布的质心位于原点，协方差矩阵为单位矩阵。图 3 的子图是在 X_3 在 5 个不同值上的“切片”，代表 $f_{X_1, X_2, X_3}(x_1, x_2, x_3 = c)$ 。容易看出来， $f_{X_1, X_2, X_3}(x_1, x_2, x_3 = c)$ 的等高线是正圆。

图 4 所示为这个三元高斯分布的边缘分布。图中我们看到了协方差矩阵分块。



本书第 12、13 章还会进一步介绍协方差矩阵分块的应用场景。

图 5 和图 6 可可视化协方差矩阵为对角阵的三元高斯分布，子图中我们看到的多是正椭圆。图 7 和图 8 可可视化协方差矩阵为一般正定矩阵的三元高斯分布，我们看到的多是旋转椭圆。

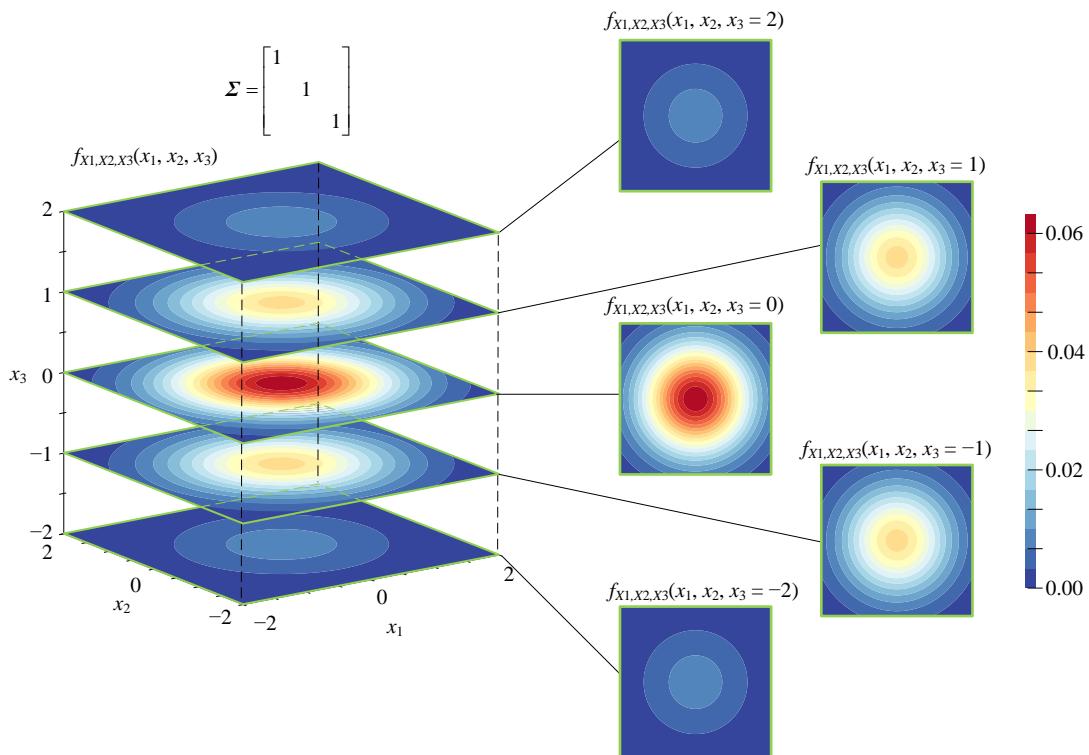


图 3. 三元高斯分布切片，协方差为单位矩阵

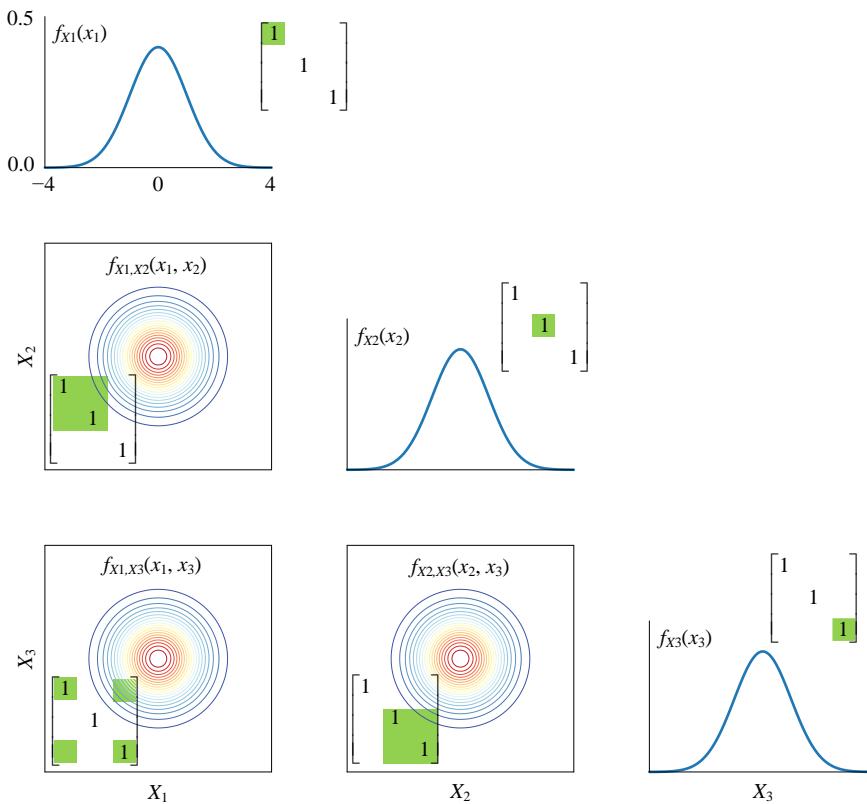


图 4. 三元高斯分布的边缘分布，协方差为单位矩阵

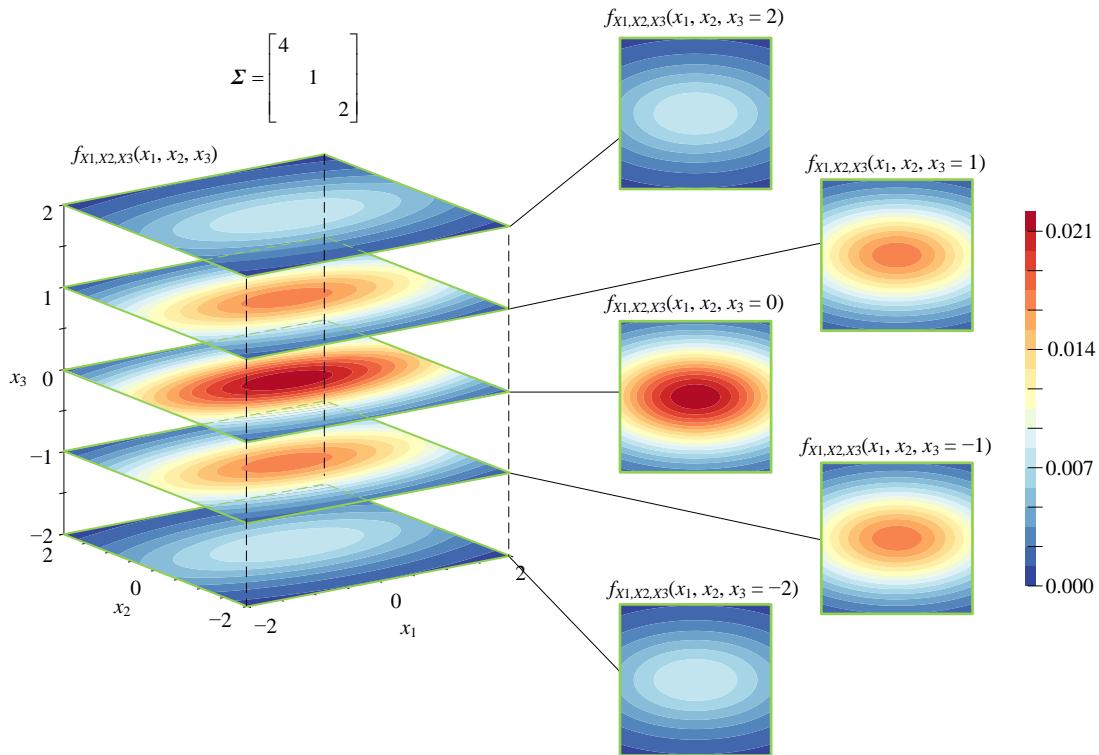


图 5. 三元高斯分布切片，协方差矩阵为对角矩阵

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

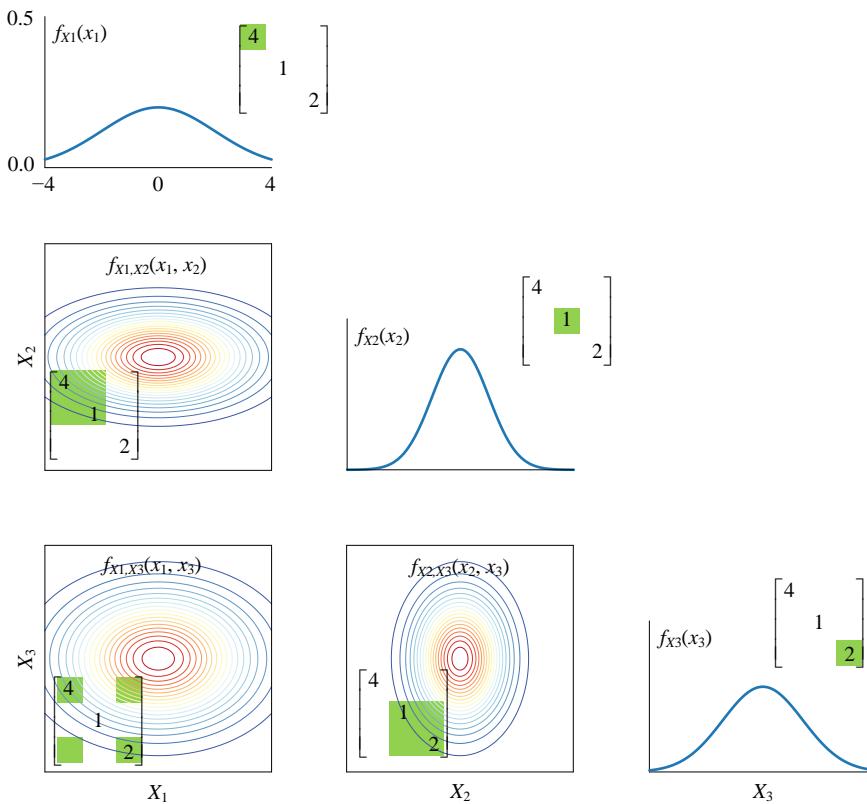


图 6. 三元高斯分布的边缘分布，协方差矩阵为对角矩阵

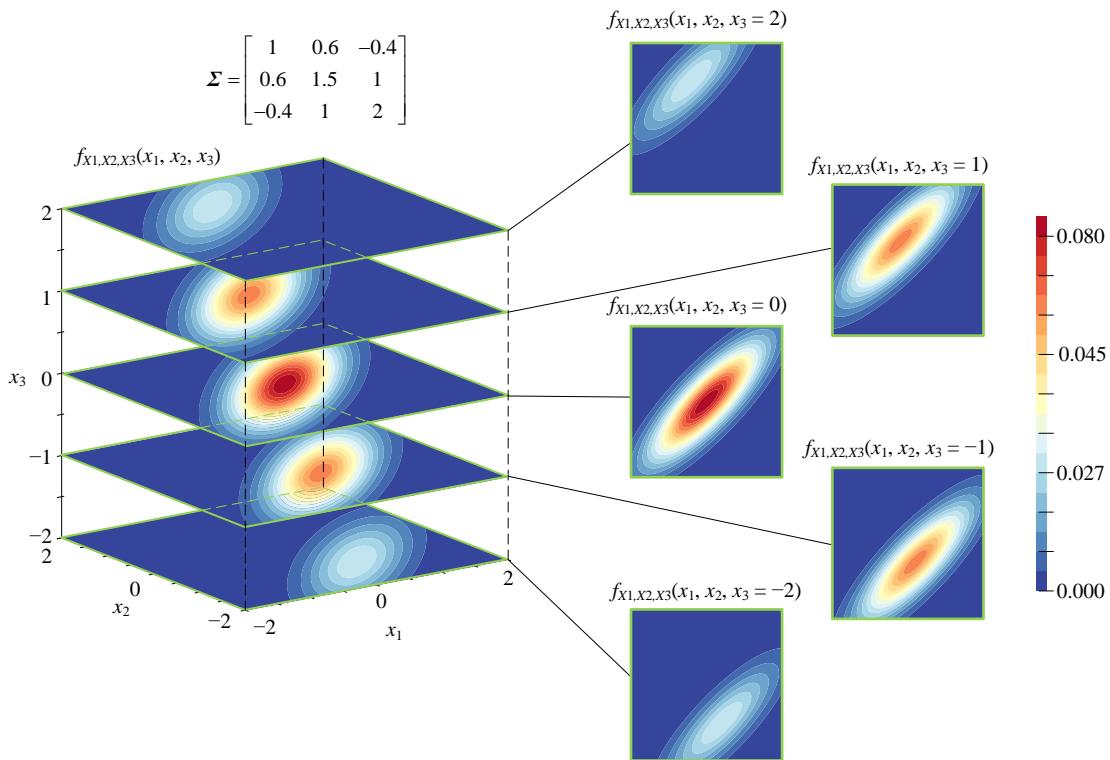


图 7. 三元高斯分布切片，协方差矩阵正定

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

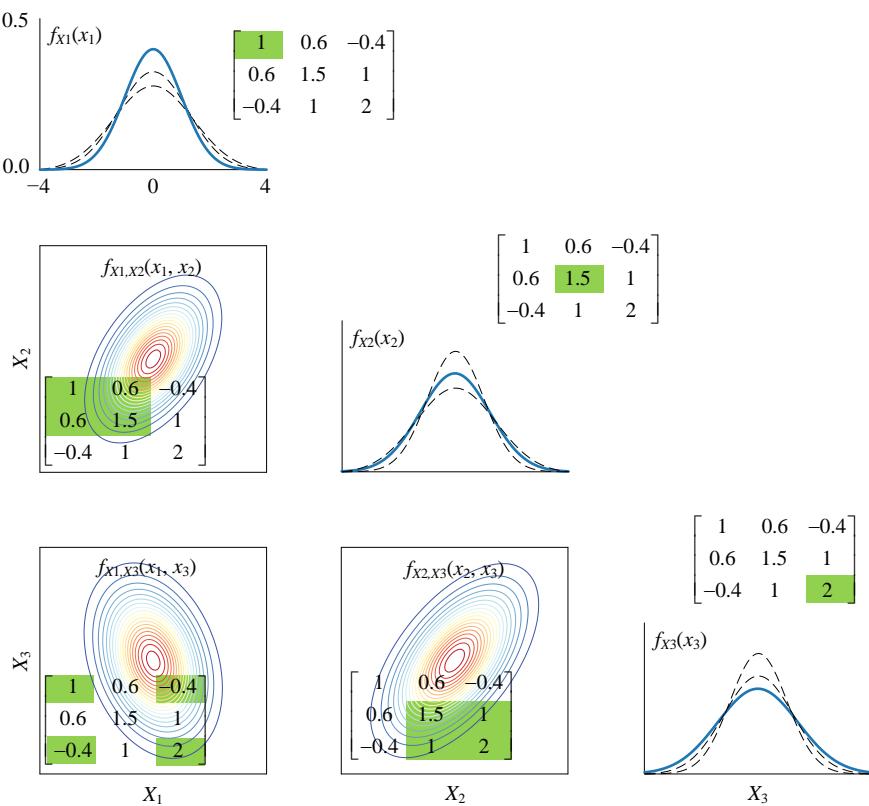


图 8. 三元高斯分布的边缘分布，协方差矩阵正定



这一章，我们用 `plotly.graph_objects.Volume()` 可视化三维高斯分布，这个 App 中大家可以调整分布参数。请大家参考 `Streamlit_Bk5_Ch11_03.py`。

11.3 解剖多元高斯分布 PDF

《矩阵力量》第 20 章介绍过如何用“平移 → 旋转 → 缩放”解剖多元高斯分布，本节把其中重要的内容“抄”了过来。

特征值分解协方差矩阵

协方差矩阵 Σ 为对称矩阵，对 Σ 谱分解得到：

$$\Sigma = V \Lambda V^T \quad (22)$$

其中， V 为正交矩阵，即满足 $V^T V = V V^T = I$ 。

如果 Σ 正定，利用(22)获得 Σ^{-1} 的特征值分解：

$$\Sigma^{-1} = V A^{-1} V^T \quad (23)$$

由此，将 $(x - \mu)^T \Sigma^{-1} (x - \mu)$ 拆成 $A^{\frac{-1}{2}} V^T (x - \mu)$ 的“平方”：

$$(x - \mu)^T V A^{-1} V^T (x - \mu) = \left[A^{\frac{-1}{2}} V^T (x - \mu) \right]^T A^{\frac{-1}{2}} V^T (x - \mu) = \left\| A^{\frac{-1}{2}} V^T (x - \mu) \right\|_2^2 \quad (24)$$

平移 → 旋转 → 缩放

(24) 的几何解释是，旋转椭圆通过“平移 $(x - \mu) \rightarrow$ 旋转 (V^T) \rightarrow 缩放 ($A^{\frac{-1}{2}}$)”转换成单位圆，具体过程如图9所示。

图9(a)中旋转椭圆代表多元高斯分布 $N(\mu, \Sigma)$ ，随机数质心位于 μ ，椭圆形形状描述了协方差矩阵 Σ 。图9(a)中散点是服从 $N(\mu, \Sigma)$ 的随机数。

图9(a)中散点经过平移得到 $x_c = x - \mu$ ，这是一个去均值(中心化过程)。图9(b)中旋转椭圆代表多元高斯分布 $N(\theta, \Sigma)$ 。随机数质心也随之平移到原点。

图9(b)中椭圆旋转之后得到图9(c)中正椭圆，对应：

$$y = V^T x_c = V^T (x - \mu) \quad (25)$$

协方差矩阵 Σ 通过特征值分解得到特征值矩阵 A 。而正椭圆的半长轴、半短轴长度蕴含在特征值矩阵 A 中，这算是拨开云雾的过程。图9(c)中随机数服从 $N(\theta, A)$ 。

最后一步是缩放，从图9(c)到图9(d)，对应：

$$z = A^{\frac{-1}{2}} y = A^{\frac{-1}{2}} V^T (x - \mu) \quad (26)$$

图9(d)中单位圆则代表多元标准分布 $N(\theta, I)$ 。这意味着满足 $N(\theta, I)$ 的随机变量为独立同分布。**独立同分布** (Independent and identically distributed, IID) 是指一组随机变量中每个变量的概率分布都相同，且这些随机变量互相独立。

利用向量 z ，多元高斯分布 PDF 可以写成：

$$f_z(z) = \frac{\exp\left(-\frac{1}{2} z^T z\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} = \frac{\exp\left(-\frac{1}{2} \|z\|_2^2\right)}{(2\pi)^{\frac{D}{2}} |A|^{\frac{1}{2}}} \quad (27)$$

z 的模 $\|z\|$ 实际上代表“整体”Z 分数。

缩放 → 旋转 → 平移

反向来看， $\mathbf{x} = \mathbf{V}\mathbf{A}^{\frac{1}{2}}\mathbf{z} + \boldsymbol{\mu}$ 代表通过“缩放 → 旋转 → 平移”把单位圆转换成中心在 $\boldsymbol{\mu}$ 的旋转椭圆。也就是把 $N(\mathbf{0}, \mathbf{I})$ 转换成 $N(\boldsymbol{\mu}, \Sigma)$ 。从数据角度来看，我们可以通过“缩放 → 旋转 → 平移”，把服从 $N(\mathbf{0}, \mathbf{I})$ 的随机数转化为服从 $N(\boldsymbol{\mu}, \Sigma)$ 的随机数。

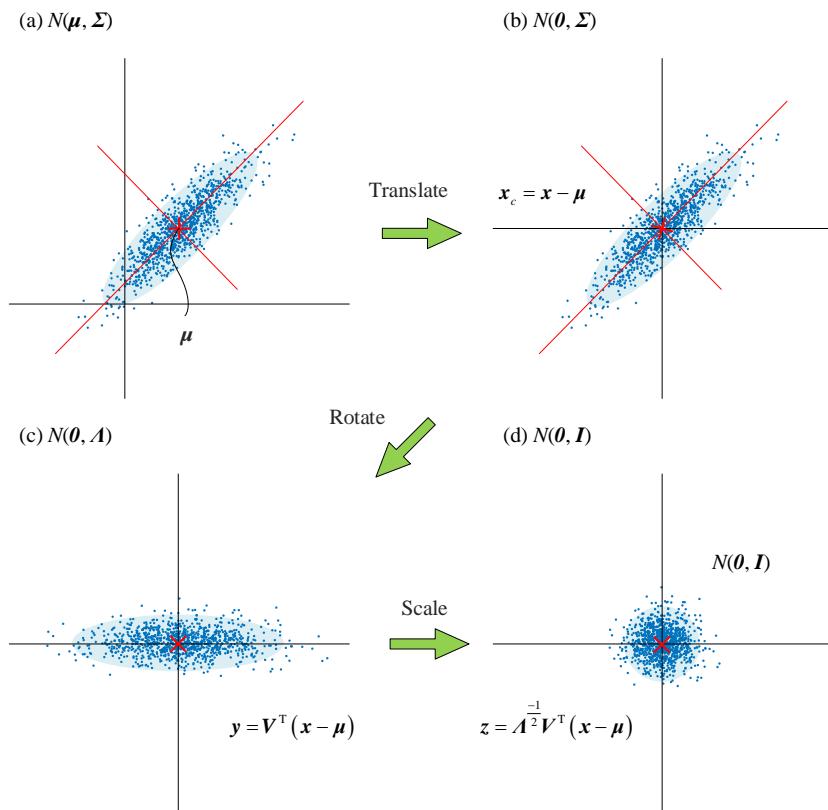


图 9. 平移 → 旋转 → 缩放，图片来自《矩阵力量》

马氏距离

马氏距离可以写成：

$$d = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})} = \left\| \mathbf{A}^{-\frac{1}{2}} \mathbf{V}^T (\mathbf{x} - \boldsymbol{\mu}) \right\| = \|z\| \quad (28)$$

马氏距离的独特之处在于，它通过引入协方差矩阵在计算距离时考虑了数据的分布。此外，马氏距离无量纲量 (unitless 或 dimensionless)，它将各个特征数据标准化。本书第 23 章将专门讲解马氏距离及其应用。

高斯函数

将(28)中马氏距离 d 代入多元高斯分布概率密度函数，得到：

$$f_{\chi}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}d^2\right)}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \quad (29)$$

上式，我们看到高斯函数 $\exp(-1/2 \bullet)$ 把“距离度量”转化成“亲近度”。图10所示为马氏距离图像。大家可以发现这个曲面为开口朝上的锥面，等高线为旋转椭圆。

图10(b)中白色虚线正圆代表距离质心 μ 欧氏距离为1的等高线。欧氏距离是最自然的距离度量。而马氏距离则引入协方差矩阵 Σ ，计算距离时考虑数据的分布情况。

 本书第23章将区分欧氏距离和马氏距离。

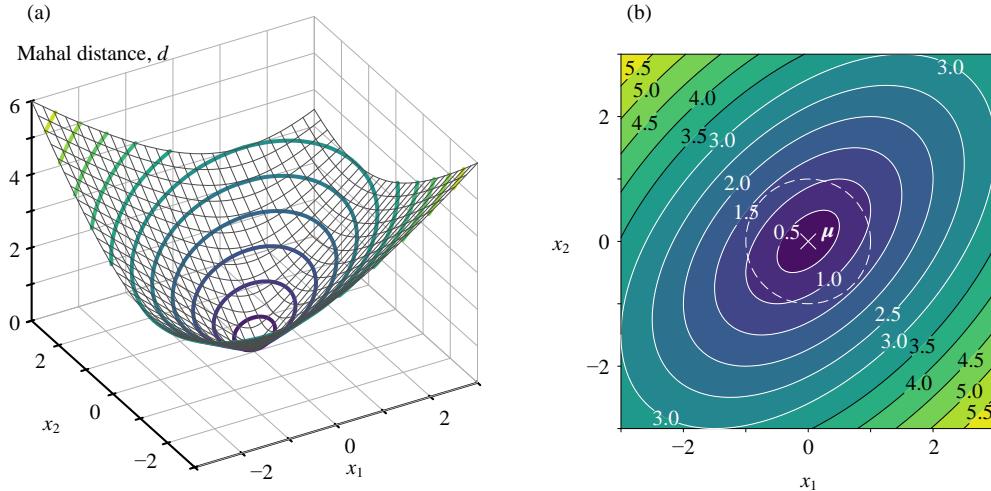


图10. 马氏距离椭圆等高线

将具体马氏距离 d 值代入(29)，可以得到高斯概率密度值。也就是说，图10每一个椭圆都对应一个概率密度值。这就是图11中等高线的含义。

 请大家注意区分，椭圆等高线到底是代表马氏距离，还是概率密度值。

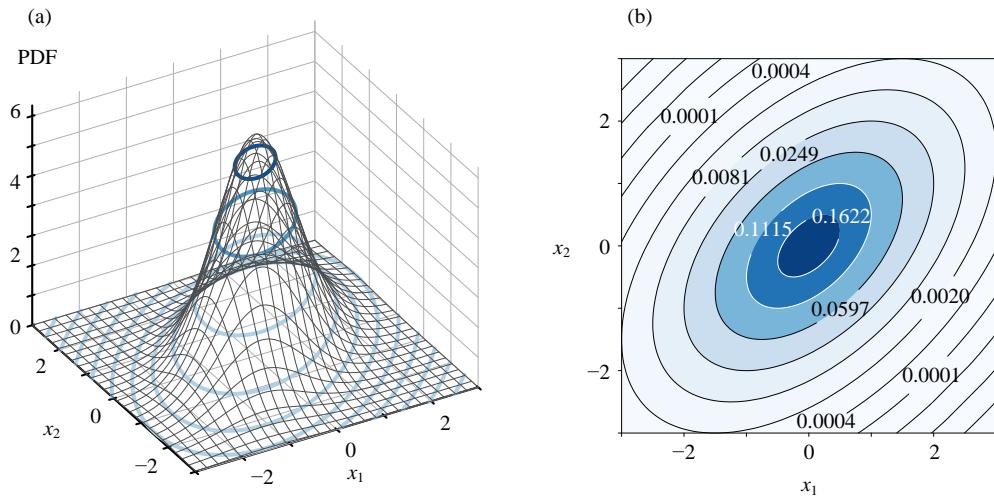


图 11. 高斯分布 PDF 椭圆等高线

分母：行列式值

把 $|\Sigma|^{\frac{1}{2}}$ 从(5) 分母移到分子可以写成 $|\Sigma|^{-\frac{1}{2}}$ 。而 $\Sigma^{\frac{-1}{2}}$ 相当于：

$$\Sigma^{\frac{-1}{2}} \sim A^{\frac{-1}{2}} V^T (x - \mu) \quad (30)$$

从体积角度来看，“平移 \rightarrow 旋转 \rightarrow 缩放”几何变换带来的面积/体积缩放系数便是 $|\Sigma|^{\frac{-1}{2}}$ 。准确来说，只有“缩放”才影响面积/体积，因此 $|\Sigma|^{\frac{-1}{2}} = |A|^{\frac{-1}{2}}$ 。

分母：体积归一化

从几何角度来看，(5) 分母中 $(2\pi)^{\frac{D}{2}}$ 一项起到归一化作用，为了保证概率密度函数曲面和整个水平面包裹的体积为 1，即概率为 1。

11.4 平移 \rightarrow 旋转

本节以二元高斯分布 PDF 为例，利用特征值分解这个工具进一步深入理解多元高斯分布。

特征值分解

形状为 2×2 协方差矩阵 Σ ，它的特征值和特征向量关系为：

$$\begin{cases} \Sigma v_1 = \lambda_1 v_1 \\ \Sigma v_2 = \lambda_2 v_2 \end{cases} \quad (31)$$

(31) 可以写成：

$$\Sigma \underbrace{\begin{bmatrix} v_1 & v_2 \end{bmatrix}}_{V} = \underbrace{\begin{bmatrix} v_1 & v_2 \end{bmatrix}}_{V} \underbrace{\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}}_A \quad (32)$$

即，

$$\Sigma V = V A \quad (33)$$

将 Σ 具体值代入 (31) 得到：两个特征值对应的特征向量如下：

$$\begin{aligned} \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} v_1 &= \lambda_1 v_1 \\ \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} v_2 &= \lambda_2 v_2 \end{aligned} \quad (34)$$

两个特征值可以通过下式求得：

$$\begin{aligned} \lambda_1 &= \frac{\sigma_1^2 + \sigma_2^2}{2} + \sqrt{\left(\rho_{1,2}\sigma_1\sigma_2\right)^2 + \left(\frac{\sigma_1^2 - \sigma_2^2}{2}\right)^2} \\ \lambda_2 &= \frac{\sigma_1^2 + \sigma_2^2}{2} - \sqrt{\left(\rho_{1,2}\sigma_1\sigma_2\right)^2 + \left(\frac{\sigma_1^2 - \sigma_2^2}{2}\right)^2} \end{aligned} \quad (35)$$

当 $\rho_{1,2} = 0$ 且 $\sigma_1 = \sigma_2$ 时，(35) 中两个特征值相等。这种条件下，概率密度等高线为正圆。

长轴、短轴

大家已经清楚，二元高斯分布的 PDF 平面等高线是椭圆。如图 12 所示， $\sqrt{\lambda_1}$ 就是椭圆半长轴长度， $\sqrt{\lambda_2}$ 就是半短轴长度：

$$\begin{aligned} EO = GO &= \sqrt{\lambda_1} = \sqrt{\frac{\sigma_x^2 + \sigma_y^2}{2} + \sqrt{\left(\rho_{x,y}\sigma_x\sigma_y\right)^2 + \left(\frac{\sigma_x^2 - \sigma_y^2}{2}\right)^2}} \\ FO = HO &= \sqrt{\lambda_2} = \sqrt{\frac{\sigma_x^2 + \sigma_y^2}{2} - \sqrt{\left(\rho_{x,y}\sigma_x\sigma_y\right)^2 + \left(\frac{\sigma_x^2 - \sigma_y^2}{2}\right)^2}} \end{aligned} \quad (36)$$

v_1 和 v_2 具体值为：

$$\mathbf{v}_1 = \begin{bmatrix} \frac{\sigma_1^2 - \sigma_2^2}{2} + \sqrt{\left(\rho_{1,2}\sigma_1\sigma_2\right)^2 + \left(\frac{\sigma_1^2 - \sigma_2^2}{2}\right)^2} \\ \rho_{1,2}\sigma_1\sigma_2 \\ 1 \end{bmatrix} \quad (37)$$

$$\mathbf{v}_2 = \begin{bmatrix} \frac{\sigma_1^2 - \sigma_2^2}{2} - \sqrt{\left(\rho_{1,2}\sigma_1\sigma_2\right)^2 + \left(\frac{\sigma_1^2 - \sigma_2^2}{2}\right)^2} \\ \rho_{1,2}\sigma_1\sigma_2 \\ 1 \end{bmatrix}$$

图 12 中， \mathbf{v}_1 对应的就是椭圆半长轴方向， \mathbf{v}_2 对应半短轴方向。在主成分分析中， \mathbf{v}_1 就是第一主元方向。 \mathbf{v}_2 便是第二主元方向。

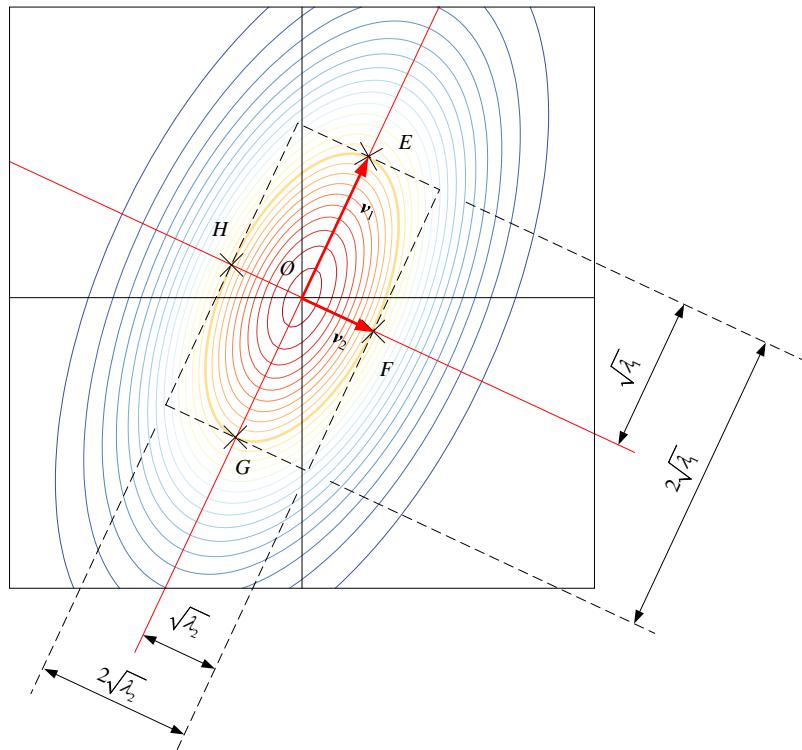


图 12. 椭圆的长轴、短轴

实际上，将 (X_1, X_2) 投影到 \mathbf{v}_1 得到的随机变量的方差就是 λ_1 ，对应的标准差为 $\sqrt{\lambda_1}$ 。将 (X_1, X_2) 投影到 \mathbf{v}_2 得到的随机变量的方差为 λ_2 ，其标准差为 $\sqrt{\lambda_2}$ 。

随机变量的线性变换

从另外一个角度来看，如图 13 所示，某个满足二元高斯分布随机变量 (X_1, X_2) 朝若干方向投影。我们先给出结论，这些方向中，向 v_1 投影得到的随机变量方差最大，向 v_2 投影得到的随机变量方差最小。

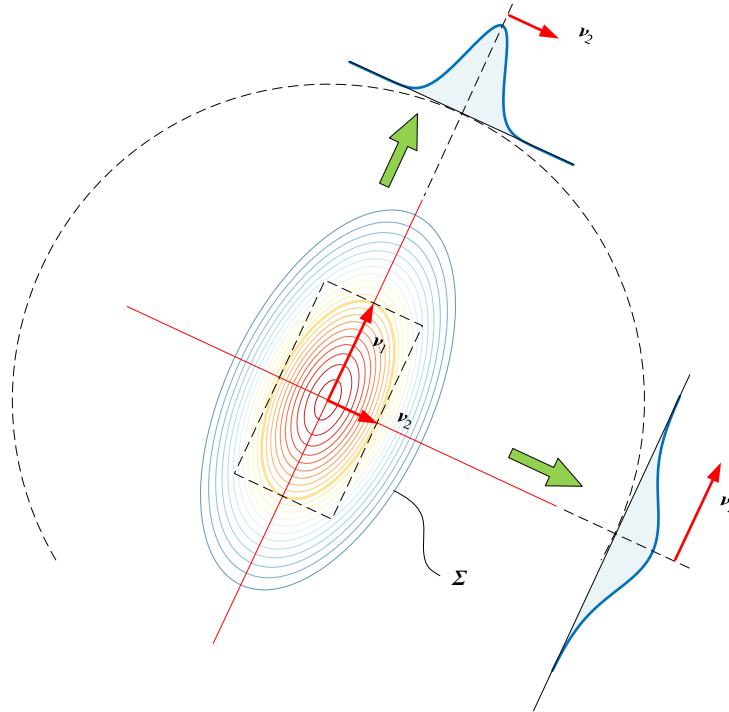


图 13. 二元高斯分布朝不同方向投影

假设二元随机变量列向量 $\chi = [X_1, X_2]^T$ 满足图 13 所示二元高斯分布。而 χ 先中心化，再向 v_1 投影得到 Y_1 ：

$$Y_1 = (\chi - \mu_\chi)^T v_1 = \left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \right)^T \begin{bmatrix} v_{1,1} \\ v_{2,1} \end{bmatrix} = (X_1 - \mu_1)v_{1,1} + (X_2 - \mu_2)v_{2,1} \quad (38)$$

从数据角度，上述过程如图 14 所示。

对 Y_1 求方差：

$$\begin{aligned} \text{var}(Y_1) &= E[(Y_1 - \mu_{Y_1})^2] = E\left[\left((\chi - \mu_\chi)^T v_1\right)^T (\chi - \mu_\chi)^T v_1\right] \\ &= v_1^T E\left[\left((\chi - \mu_\chi)^T\right)\left(\chi - \mu_\chi\right)^T\right] v_1 \\ &= v_1^T \Sigma_\chi v_1 \end{aligned} \quad (39)$$

因为 Y_1 已经中心化，所以上式中 $\mu_{Y_1} = 0$ 。

将 Σ_x 的特征值分解代入 (39) 得到：

$$\begin{aligned}\text{var}(Y_1) &= \mathbf{v}_1^T \Sigma_x \mathbf{v}_1 = \mathbf{v}_1^T [\mathbf{v}_1 \quad \mathbf{v}_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix} \mathbf{v}_1 \\ &= [1 \quad 0] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \lambda_1\end{aligned}\quad (40)$$

实际上就是随机变量的线性变换，我们将会在本书第 14 章继续这一话题。

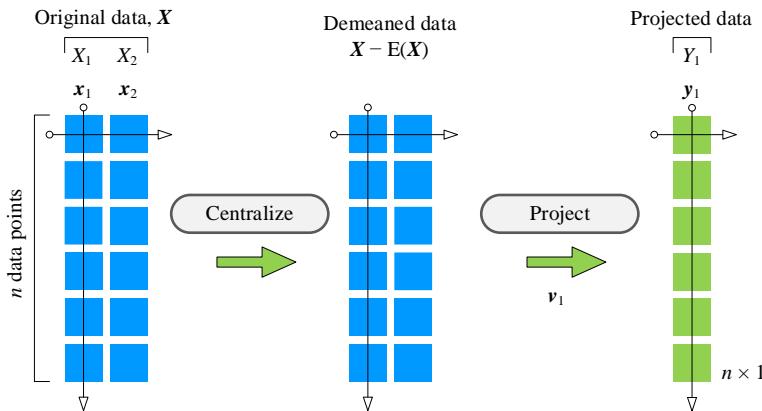


图 14. X 先中心化，再向 v_1 投影得到 y_1

椭圆旋转

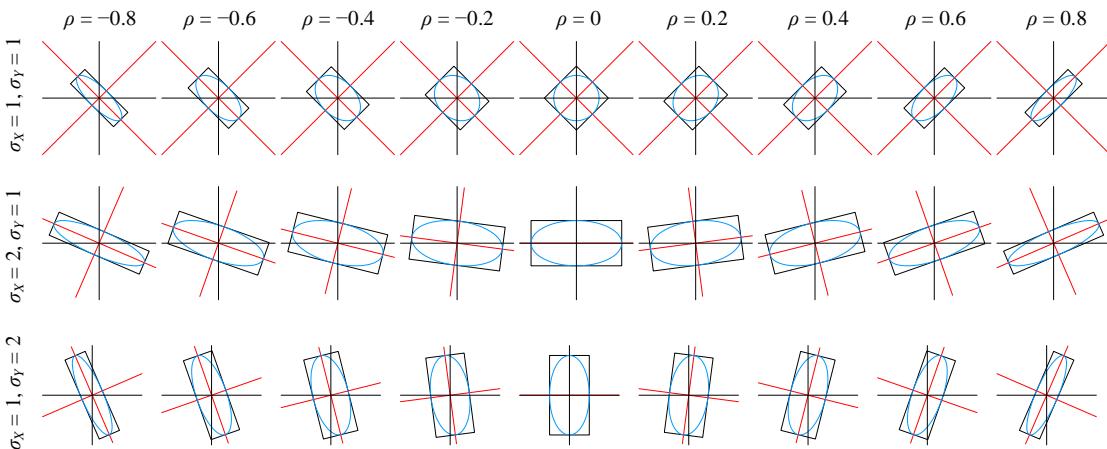
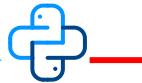
椭圆旋转角度 θ :

$$\theta = \frac{1}{2} \arctan \left(\frac{2\rho_{1,2}\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2} \right) \quad (41)$$

图 15 所示为在 σ_1, σ_2 大小不同， $\rho_{1,2}$ 取值不同对椭圆旋转的影响。

观察 (41)，发现椭圆的旋转角度和 $\sigma_1, \sigma_2, \rho_{1,2}$ 有关。

特别地，当 $\sigma_1 = \sigma_2$ 时，如果 $\rho_{1,2}$ 为小于 1 的正数，椭圆的旋转角度为 45° ；如果 $\rho_{1,2}$ 为大于 -1 的负数，椭圆的旋转角度为 -45° 。

图 15. 在 σ_1 、 σ_2 大小不同， $\rho_{1,2}$ 取值不同对椭圆旋转的影响

Bk5_Ch11_02.py 绘制图 15。

特征值之和

可以发现 (35) 中两个特征值之和，等于协方差矩阵 Σ 的两个方差之和：

$$\lambda_1 + \lambda_2 = \sigma_1^2 + \sigma_2^2 \quad (42)$$



这正是《矩阵力量》讲到的特征值分解中，原矩阵迹等于特征值矩阵的迹。建议大家回顾特征值分解的优化视角。

特征值之积

两个特征值乘积为：

$$\begin{aligned} \lambda_1 \lambda_2 &= \left(\frac{\sigma_1^2 + \sigma_2^2}{2} \right)^2 - \left((\rho_{1,2} \sigma_1 \sigma_2)^2 + \left(\frac{\sigma_1^2 - \sigma_2^2}{2} \right)^2 \right) \\ &= \sigma_1^2 \sigma_2^2 - \rho_{1,2}^2 \sigma_1^2 \sigma_2^2 = \sigma_1^2 \sigma_2^2 (1 - \rho_{1,2}^2) \end{aligned} \quad (43)$$

这和协方差矩阵 Σ 行列式值相等：

$$|\Sigma| = \sigma_1^2 \sigma_2^2 - \rho_{1,2}^2 \sigma_1^2 \sigma_2^2 = \sigma_1^2 \sigma_2^2 (1 - \rho_{1,2}^2) \quad (44)$$

谱分解

Σ 的谱分解可以进一步写成：

$$\Sigma = V \Lambda V^T = [\mathbf{v}_1 \quad \mathbf{v}_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{bmatrix} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T \quad (45)$$

本书下一章还会继续这一话题。

平移 \rightarrow 旋转

令

$$\mathbf{y} = V^T (\mathbf{x} - \boldsymbol{\mu}) \quad (46)$$

发现上式 $V^T(\mathbf{x} - \boldsymbol{\mu})$ 相当于 \mathbf{x} 经过平移 $(\mathbf{x} - \boldsymbol{\mu})$ 、旋转 (V^T) 两步操作得到 \mathbf{y} 。整个过程如图 16 所示。

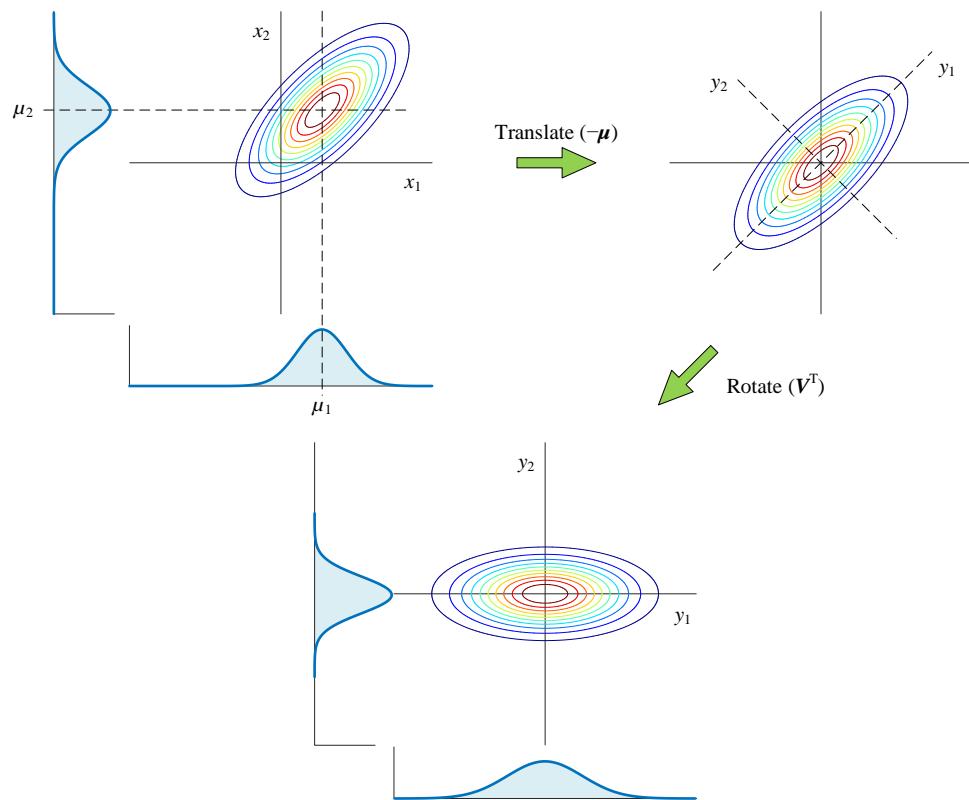


图 16. 椭圆先平移再旋转

这样 $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ 可以写成：

$$\mathbf{y}^T \boldsymbol{\Lambda}^{-1} \mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_q \end{bmatrix} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_q \end{bmatrix}^{-1} \begin{bmatrix} y_1 & y_2 & \cdots & y_q \end{bmatrix}^T = \sum_{j=1}^D \frac{y_j^2}{\lambda_j} \quad (47)$$

其中， $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$ 。上式代表着一个多维空间正椭球体。

平移 $(\mathbf{x} - \boldsymbol{\mu})$ 、旋转 (\mathbf{V}^T) 两步几何变换只改变椭球的空间位置和旋转角度，不改变椭球本身的几何尺寸。也就是说， $|\boldsymbol{\Sigma}| = |\boldsymbol{\Lambda}|$ 。

特别地，当 $D = 2$ 时，令 $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ 为 1，(47) 可以写成平面正椭圆：

$$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = 1 \quad (48)$$

显然这个椭圆中心位于原点，同样这就解释了为什么图 12 中椭圆的半长轴为 $\sqrt{\lambda_1}$ ，半短轴为 $\sqrt{\lambda_2}$ 。

反过来， \mathbf{y} 先经过旋转、再平移得到 \mathbf{x} ：

$$\mathbf{x} = \mathbf{V}\mathbf{y} + \boldsymbol{\mu} \quad (49)$$

独立

二元随机变量 (Y_1, Y_2) 对应的二元高斯分布 PDF 为：

$$\begin{aligned} f_{Y_1, Y_2}(y_1, y_2) &= \frac{1}{2\pi\sqrt{\lambda_1\lambda_2}} \times \exp\left(-\frac{1}{2}\left(\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}\right)\right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}\sqrt{\lambda_1}} \exp\left(-\frac{1}{2}\frac{y_1^2}{\lambda_1}\right)}_{f_{Y_1}(y_1)} \times \underbrace{\frac{1}{\sqrt{2\pi}\sqrt{\lambda_2}} \exp\left(-\frac{1}{2}\frac{y_2^2}{\lambda_2}\right)}_{f_{Y_2}(y_2)} \end{aligned} \quad (50)$$

可以发现随机变量 Y_1 和 Y_2 独立。如图 16 所示，随机变量 Y_1 对应的方差为 λ_1 ，标准差为 $\sqrt{\lambda_1}$ ；随机变量 Y_2 对应的方差为 λ_2 ，标准差为 $\sqrt{\lambda_2}$ 。

11.5 平移 → 旋转 → 缩放

$(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ 可以整理为：

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = [V^T (x - \mu)]^T A^{\frac{-1}{2}} A^{\frac{-1}{2}} [V^T (x - \mu)] = \left(A^{\frac{-1}{2}} V^T (x - \mu) \right)^2 \quad (51)$$

这就是前文讲到的“开方”。

令：

$$z = A^{\frac{-1}{2}} V^T (x - \mu) \quad (52)$$

上式相当于 x 经过平移、旋转和缩放，最后得到 z ，整个过程如图 17 所示。

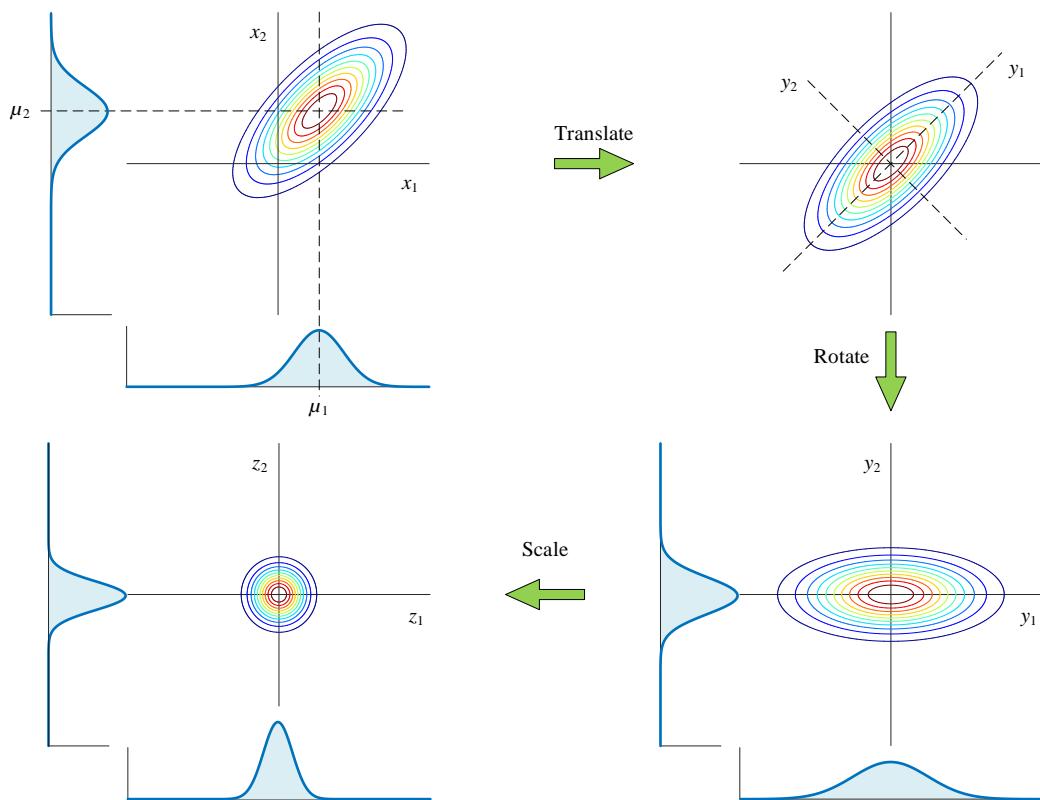


图 17. 椭圆先平移、再旋转，最后缩放，得到单位圆

单位球体

将 (52) 代入 (51)，得到的解析式：

$$(x - \mu)^T \Sigma^{-1} (x - \mu) = z^T z = z_1^2 + z_2^2 + \dots + z_D^2 = \sum_{j=1}^D z_j^2 \quad (53)$$

当上式为 1 时，它代表多维空间的单位球体。

反过来，也可以利用 z 通过缩放、旋转、平移，反求 x ：

$$\mathbf{x} = \mathbf{V} \mathbf{D} \mathbf{z} + \boldsymbol{\mu} \quad (54)$$

Rotate Scale Translate

图 18 展示 (54) 对应的几何变换。

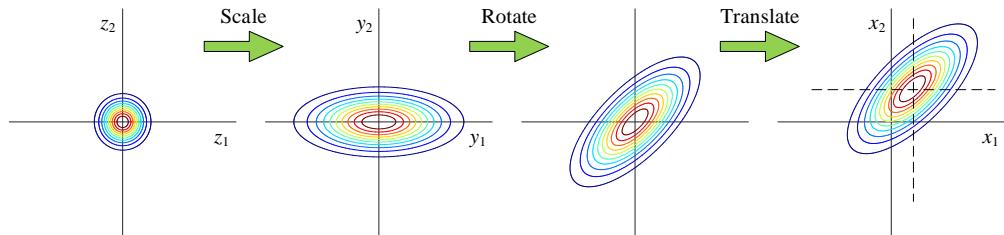


图 18. 单位圆先缩放，再旋转，最后平移

数据视角

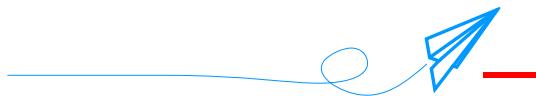
类似 (38)，从数据角度来看，如果数据矩阵 \mathbf{X} 服从 $N(\mathbf{E}(\mathbf{X}), \Sigma_{\mathbf{X}})$ 。对 \mathbf{X} 先中心化，再向 \mathbf{V} 投影，最后缩放得到 \mathbf{Z} ：

$$\mathbf{Z} = (\mathbf{X} - \mathbf{E}(\mathbf{X})) \mathbf{V} \mathbf{A}^{-\frac{1}{2}} \quad (55)$$

\mathbf{Z} 的协方差矩阵为单位矩阵 \mathbf{I} ：

$$\begin{aligned} \Sigma_{\mathbf{Z}} &= \mathbf{Z}^T \mathbf{Z} = \frac{\left((\mathbf{X} - \mathbf{E}(\mathbf{X})) \mathbf{V} \mathbf{A}^{-\frac{1}{2}} \right)^T \left((\mathbf{X} - \mathbf{E}(\mathbf{X})) \mathbf{V} \mathbf{A}^{-\frac{1}{2}} \right)}{n-1} \\ &= \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^T \frac{\overbrace{(\mathbf{X} - \mathbf{E}(\mathbf{X}))^T (\mathbf{X} - \mathbf{E}(\mathbf{X}))}^{\Sigma_{\mathbf{X}}}}{n-1} \mathbf{V} \mathbf{A}^{-\frac{1}{2}} \\ &= \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^T \Sigma_{\mathbf{X}} \mathbf{V} \mathbf{A}^{-\frac{1}{2}} = \mathbf{I} \end{aligned} \quad (56)$$

也就是说，如果 \mathbf{X} 服从多维高斯分布的话， \mathbf{Z} 服从 IID 标准正态分布。



本章将一元、二元、三元高斯分布提高到了多元。而多元高斯分布离不开矩阵运算。

利用特征值分解，我们从几何角度理解多元高斯分布 PDF 中隐含的“平移 → 旋转 → 缩放”。这对理解协方差矩阵、马氏距离、主成分分析等概念至关重要。

以后希望大家每次见到多元高斯分布 PDF 式子，对它的每个组成部分的作用都能如数家珍、滔滔不绝。

12

Conditional Gaussian Distributions

条件高斯分布

假设随机变量服从高斯分布，讨论条件期望、条件方差



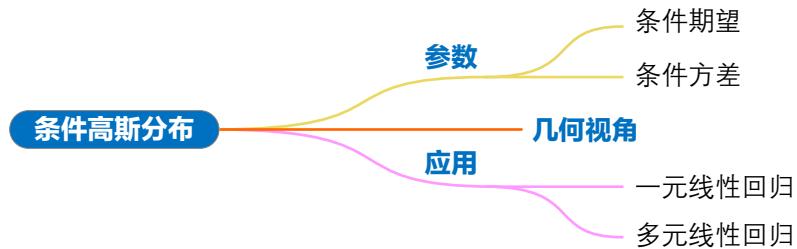
生命就像一个永恒的春天，穿着崭新而绚丽的衣服站在我面前。

Life stands before me like an eternal spring with new and brilliant clothes.

——卡尔·弗里德里希·高斯 (Carl Friedrich Gauss) | 德国数学家、物理学家、天文学家 | 1777 ~ 1855



- ◀ `matplotlib.pyplot.contour()` 绘制等高线图
- ◀ `matplotlib.pyplot.contour3D()` 绘制三维等高线图
- ◀ `matplotlib.pyplot.contourf()` 绘制填充等高线图
- ◀ `matplotlib.pyplot.fill_between()` 区域填充颜色
- ◀ `matplotlib.pyplot.plot_wireframe()` 绘制线框图
- ◀ `scipy.stats.multivariate_normal()` 多元正态分布对象
- ◀ `scipy.stats.norm()` 一元正态分布对象



12.1 联合概率和条件概率关系

本章是本书第 8 章的延续。本书第 8 章专门介绍了离散、连续随机变量的条件 **期望** (conditional expectation)、条件 **方差** (conditional variance)。本章将这些数学工具用在高斯分布上。

本节首先回顾**条件概率** (conditional probability)。

条件概率

本章第 3 章介绍过，条件概率是指某事件在另外一个事件已经发生条件下的概率。

以图 1 为例， X 和 Y 为连续随机变量， (X, Y) 服从二元高斯分布。 (X, Y) 的联合概率密度函数 PDF $f_{X,Y}(x,y)$ 为图 1 所示曲面。

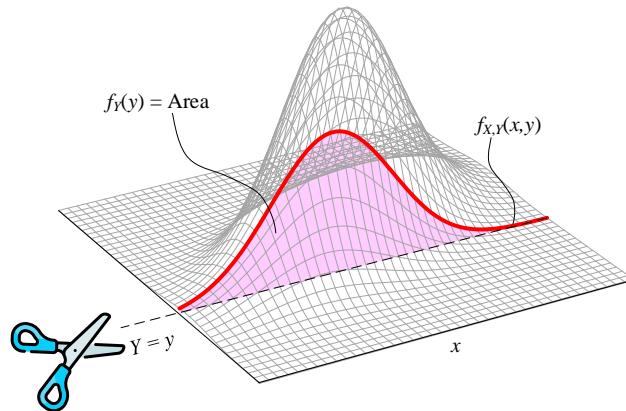


图 1. 高斯二元分布 PDF 曲面沿着 $Y=y$ 切一刀

给定 $Y=y$ 条件下，相当于在图 1 上沿着 $Y=y$ 切一刀，得到的红色曲线便是 $f_{X|Y}(x|y)$ 。

几何视角来看，给定 $Y=y$ 的条件下 ($f_Y(y) > 0$)，利用贝叶斯定理， X 的条件 PDF $f_{X|Y}(x|y)$ 相当于对 $f_{X,Y}(x, y)$ 曲线用边缘 PDF $f_Y(y)$ 归一化：

$$\overbrace{f_{X|Y}(x|y)}^{\substack{\text{Conditional} \\ \text{Given } Y=y}} = \frac{\overbrace{f_{X,Y}(x,y)}^{\text{Joint}}}{\underbrace{f_Y(y)}_{\text{Marginal}}} \quad (1)$$

⚠ 注意，此时 $f_Y(y)$ 代表一个具体的值，但是这个值仍然是概率密度，而不是概率。

分解来看， $Y=y$ 时，联合 PDF $f_{X,Y}(x,y)$ 这条曲线和横轴围成的面积为边缘 PDF $f_Y(y)$ ，即：

$$f_Y(y) = \int_x f_{X,Y}(x, y) dx$$

(2)

归一化后的 $f_{X|Y}(x|y)$ 曲线和横轴围成的面积为 1，即：

$$\int_x f_{X|Y}(x|y) dx = 1$$

(3)

沿着这个思路，让我们观察一组当 Y 取不同值时，高斯二元分布联合概率和条件概率的关系。

Y取特定值

如图 2 所示，当 $y = -2$ 时，对联合 PDF 曲面在 $y = -2$ 处切一刀，得到 $f_{X,Y}(x, y = -2)$ 对应图 2 中红色曲线。

$f_{X,Y}(x, y = -2)$ 和横轴围成的面积便是边缘 PDF $f_Y(y = -2)$ ，经过计算得知面积约为 0.05，即 $f_Y(y = -2) = 0.05$ 。

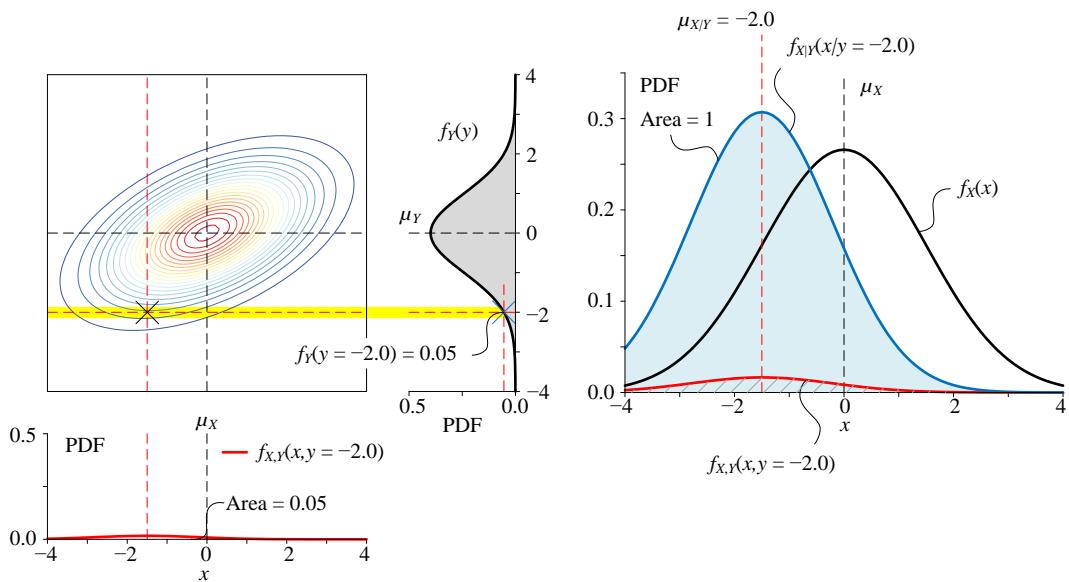
⚠ 再次强调，0.05 不是概率值，虽然它的大小某种程度上也代表“可能性”。

在给定 $y = -2$ 条件下，条件 PDF $f_{X|Y}(x|y = -2)$ 可以通过下式计算得到：

$$f_{X|Y}(x|y = -2) = \frac{f_{X,Y}(x, y = -2)}{f_Y(y = -2)}$$

(4)

图 2 同时比较联合 PDF $f_{X,Y}(x, y = -2)$ 、边缘 PDF $f_X(x)$ 、条件 PDF $f_{X|Y}(x|y = -2)$ 三条曲线之间的关系。

图 2. $y = -2$ 时，联合 PDF、边缘 PDF、条件 PDF 的关系

从图像上可以清楚看到，条件 PDF $f_{X|Y}(x|y = -2)$ 相当于联合 PDF $f_{X,Y}(x, y = -2)$ 在高度上放大约 20 倍 ($= 1/0.05$)。

⚠ 值得反复强调的是，联合 PDF $f_{X,Y}(x, y = -2)$ 曲线和横轴围成的面积约为 0.05，然而条件 PDF $f_{X|Y}(x|y = -2)$ 曲线和横轴围成的面积为 1。

Y取不同值

图 2 到图 6 五幅图分别展示当 y 取值分别为 $-2, -1, 0, 1, 2$ 时，联合 PDF 和条件 PDF 关系。

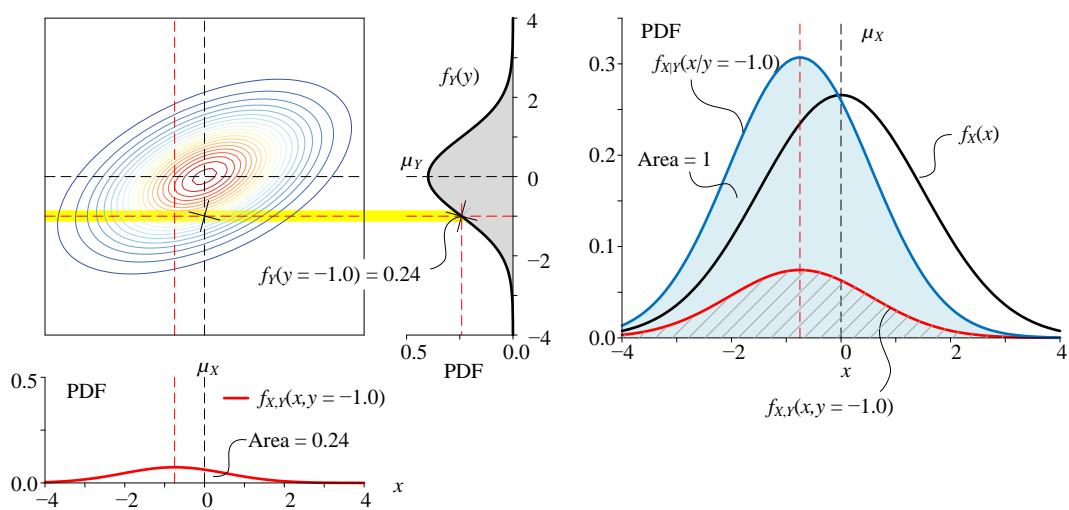
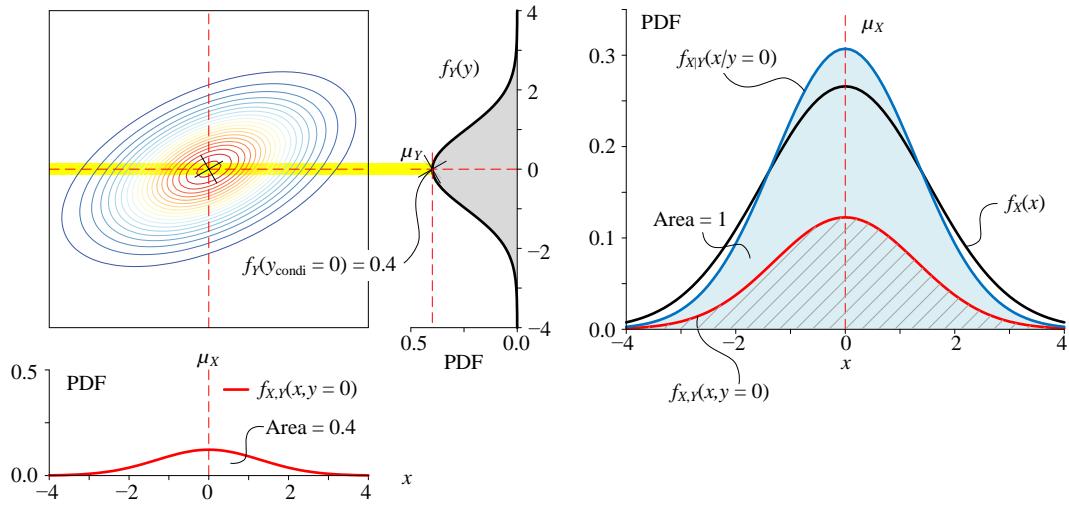
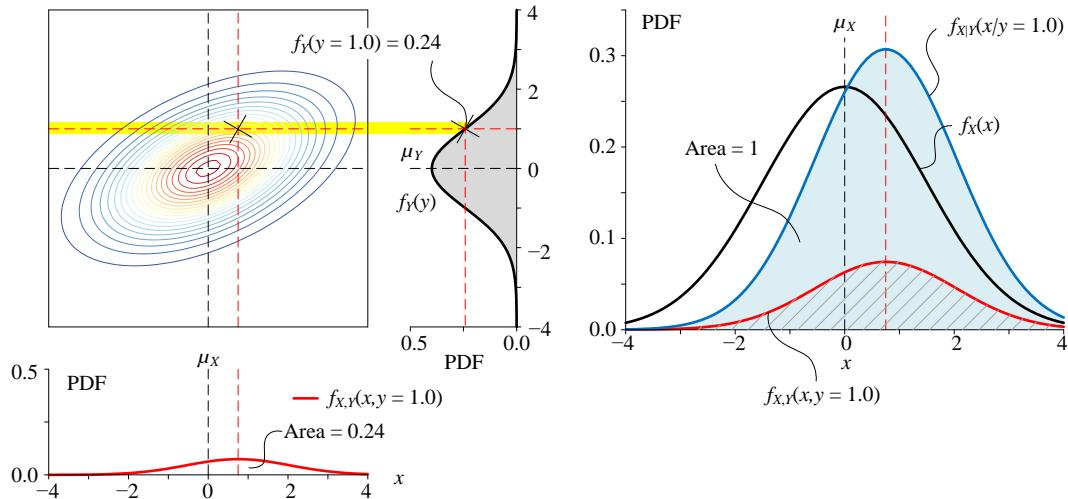
有几点值得注意。五幅图像上概率曲线形状都是类似高斯一元分布曲线。它们本身不是一元随机变量 PDF 的原因很简单——面积不为 1。经过缩放得到面积为 1 的曲线就是条件 PDF。

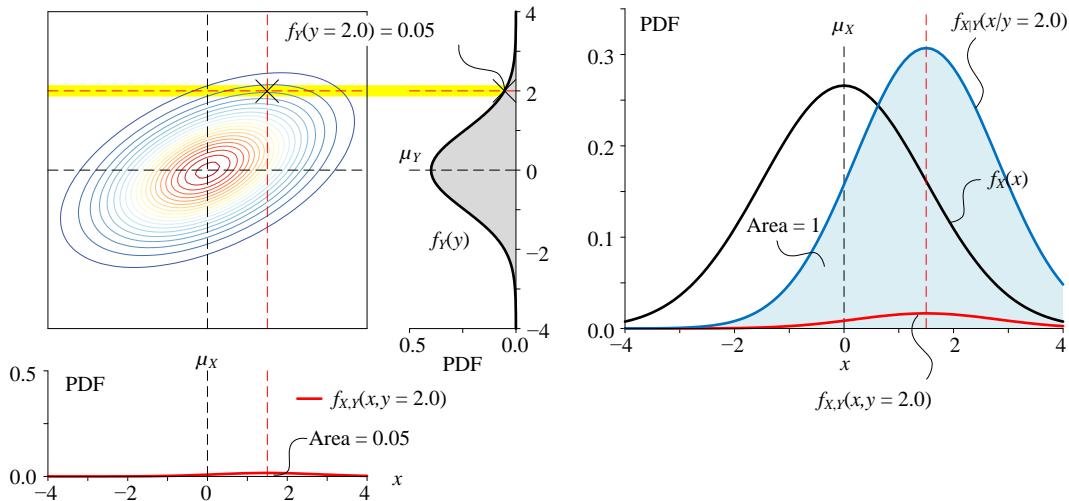
$Y = y$ 直线和联合 PDF 等高线某一个椭圆相切，而当 y 变化时，切点似乎沿着直线运动。

切点的横轴取值对应条件 PDF $f_{X|Y}(x|y)$ 曲线的对称轴，而这个对称轴又是条件 PDF $f_{X|Y}(x|y)$ 曲线的**期望**。这个**期望**值就是本书第 8 章介绍的条件**期望** (conditional expectation) $E(X|Y = y)$ 。

图 2 到图 6 五幅图条件 PDF $f_{X|Y}(x|y)$ 对应的蓝色曲线，似乎在形状上没有任何变化，仅仅是对称轴发生移动。这一点说明， y 取值变化时，条件 PDF 曲线对应分布的**方差**似乎没有变化；这个**方差**就是本书第 8 章介绍的条件**方差** (conditional variance) $\text{var}(X|Y = y)$ 。

这一节先给大家一个直观印象，本章之后将会利用高斯二元分布对条件概率、条件**期望**、条件**方差**等概念进行定量研究。

图 3. $y = -1.0$ 时，联合 PDF、边缘 PDF、条件 PDF 的关系图 4. $y = 0$ 时，联合 PDF、边缘 PDF、条件 PDF 的关系图 5. $y = 1.0$ 时，联合 PDF、边缘 PDF、条件 PDF 的关系

图 6. $y=2$ 时，联合 PDF、边缘 PDF、条件 PDF 的关系

Bk5_Ch12_01.py 绘制图 2 ~ 图 6。

12.2 给定 X 条件下， Y 的条件概率：以二元高斯为例

如果 (X, Y) 服从二元高斯分布，联合 PDF $f_{X,Y}(x,y)$ 解析式如下：

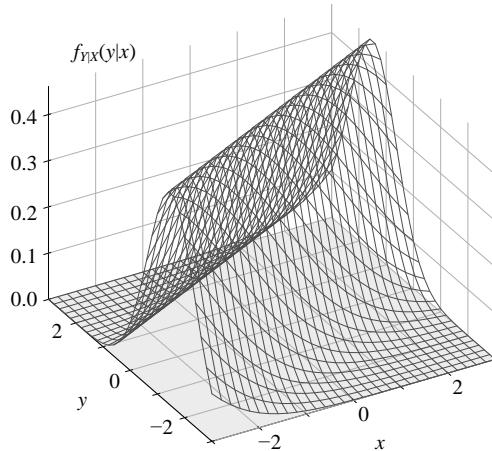
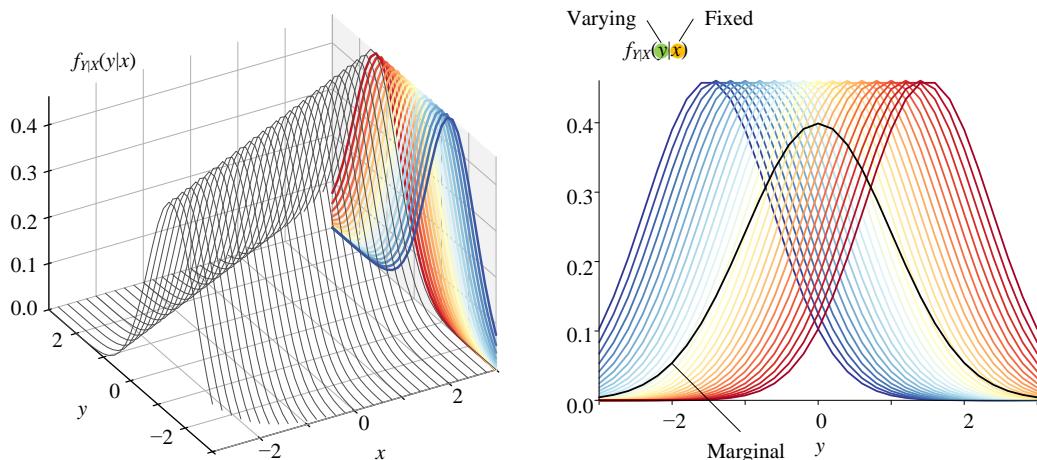
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{X,Y}^2}} \times \exp\left(-\frac{1}{2}\underbrace{\frac{1}{1-\rho_{X,Y}^2}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{X,Y}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)}_{\text{Ellipse}}\right) \quad (5)$$

利用条件 PDF、联合 PDF、边缘 PDF 三者关系，我们可以求得在给定 $X=x$ 条件下，条件 PDF $f_{Y|X}(y|x)$ 解析式为：

$$f_{Y|X}(y|x) = \frac{1}{\sigma_Y\sqrt{1-\rho_{X,Y}^2}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - \left(\mu_Y + \rho_{X,Y}\frac{\sigma_Y}{\sigma_X}(x - \mu_X)\right)}{\sigma_Y\sqrt{1-\rho_{X,Y}^2}}\right)^2\right) \quad (6)$$

图 7 所示为 $f_{Y|X}(y|x)$ 曲面网格线。 $f_{Y|X}(y|x)$ 这个曲线的**期望**和**方差**对应条件**期望** $E(Y|X=x)$ 和条件**方差** $\text{var}(Y|X=x)$ 。

可以发现当 $X = x$ 取一定值时，(6) 解析式对应高斯正态分布，这印证了本书第 10 章的猜测。将 $f_{Y|X}(y|x)$ 曲面不同位置曲线投影在 yz 平面得到图 8，容易发现这些曲线的形状完全相同（条件 **标准差** 不变），但是曲线的中心位置变化（条件 **期望值** 变化）。

图 7. $f_{Y|X}(y|x)$ 曲面网格线图 8. $f_{Y|X}(y|x)$ 曲面在 yz 平面上投影

条件期望 $E(Y|X=x)$

如果 (X, Y) 满足二元高斯分布，给定 $X=x$ 条件下， Y 的条件 PDF $f_{Y|X}(y|x)$ 如图 9 所示。图 10 所示为 $f_{Y|X}(y|x)$ 平面等高线。条件 **期望** $E(Y|X=x)$ 解析式为：

$$E(Y|X=x) = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_X) \quad (7)$$

如图 10 所示， $E(Y|X=x)$ 随着 $X=x$ 取值线性变化；也就是说， $E(Y|X=x)$ 和 x 的关系是一条直线。这条直线的一般式可以写成：

$$y = \mu_Y + \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_X) \quad (8)$$

可以发现直线的斜率为 $\rho_{X,Y}\sigma_Y/\sigma_X$, 且通过点 (μ_X, μ_Y) 。眼尖的读者一眼就会发现, 这条曲线是 x 为自变量、 y 为因变量的 OLS 线性回归直线解析式。



本章最后一节将深入探讨这一话题, 此外本书第 24 章也会展开讲解线性回归。

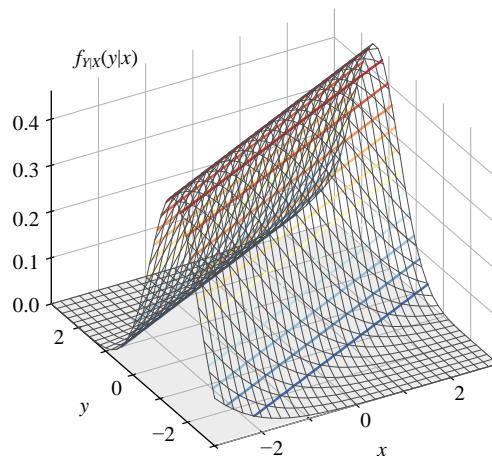


图 9. $f_{Y|X}(y|x)$ 曲面等高线

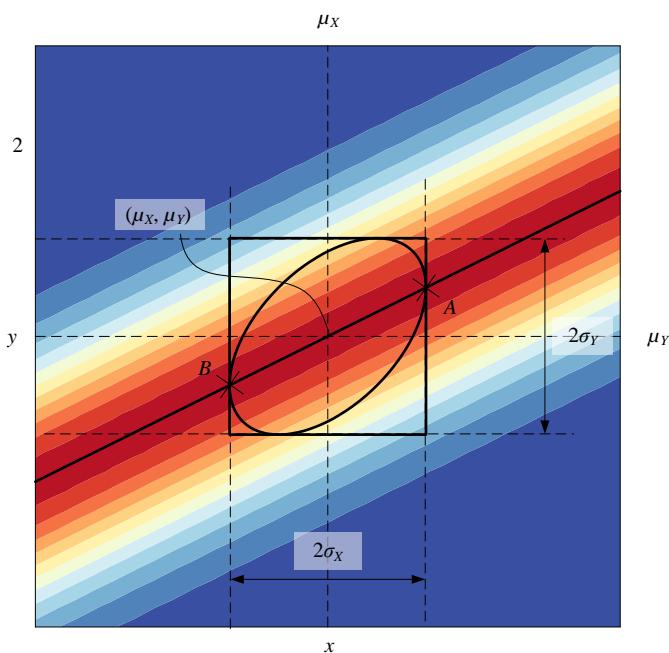


图 10. $f_{Y|X}(y|x)$ 平面等高线

条件方差 $\text{var}(Y|X = x)$

给定 $X=x$ 条件下, Y 的条件方差 $\text{var}(Y|X=x)$ 解析式为:

$$\text{var}(Y|X=x) = (1 - \rho_{X,Y}^2) \sigma_Y^2 \quad (9)$$

给定 $X=x$ 条件下， Y 的条件**标准差** $\sigma_{Y|X=x}$ 解析式为定值：

$$\sigma_{Y|X=x} = \sqrt{1 - \rho_{X,Y}^2} \cdot \sigma_Y \quad (10)$$

这解释了为什么图 10 中的等高线为平行线。

图 11 所示为 $\sigma_{Y|X=x}$ 的几何含义。



请大家格外注意图中的平行四边形，我们将在本书第 15 章还会看到这个平行四边形。

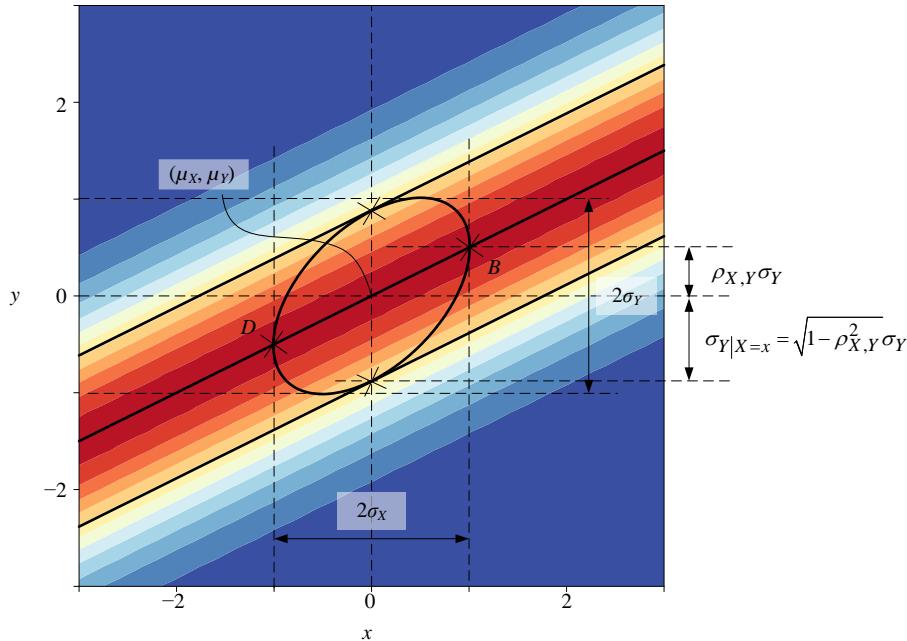
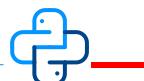


图 11. 条件**标准差** $\sigma_{Y|X}$ 的几何含义



Bk5_Ch12_02.py 绘制图 7 ~ 图 10。

以鸳尾花为例：条件期望 $E(X_2 | X_1 = x_1)$ 、**条件方差** $\text{var}(X_2 | X_1 = x_1)$

以鸳尾花花萼长度 (X_1)、花萼宽度 (X_2) 数据为例，假设 (X_1, X_2) 服从二元高斯分布。条件 PDF $f_{X_2 | X_1}(x_2 | x_1)$ 三维等高线和平面等高线如图 12 所示。

在给定 $X_1 = x_1$ 条件下， X_2 的条件**期望** $E(X_2 | X_1 = x_1)$ 解析式为：

$$\begin{aligned}
 E(X_2 | X_1 = x_1) &= \mu_2 + \rho_{1,2} \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1) \\
 &= 3.057 - 0.117 \times \frac{0.434}{0.825} (x_1 - 5.843) \\
 &= -0.615x_1 + 3.417
 \end{aligned} \tag{11}$$

条件方差 $\text{var}(X_2 | X_1 = x_1)$ 为：

$$\text{var}(X_2 | X_1 = x_1) = (1 - \rho_{1,2}^2) \sigma_2^2 \approx 0.186 \tag{12}$$

条件标准差 $\sigma_{X_2 | X_1 = x_1}$ 为：

$$\sigma_{X_2 | X_1 = x_1} = \sqrt{1 - \rho_{1,2}^2} \sigma_2 = 0.431 \tag{13}$$

如图 12 所示，不管 x_1 怎么变，这个条件标准差 $\sigma_{X_2 | X_1 = x_1}$ 为定值。请大家对比第 8 章的类似图片。

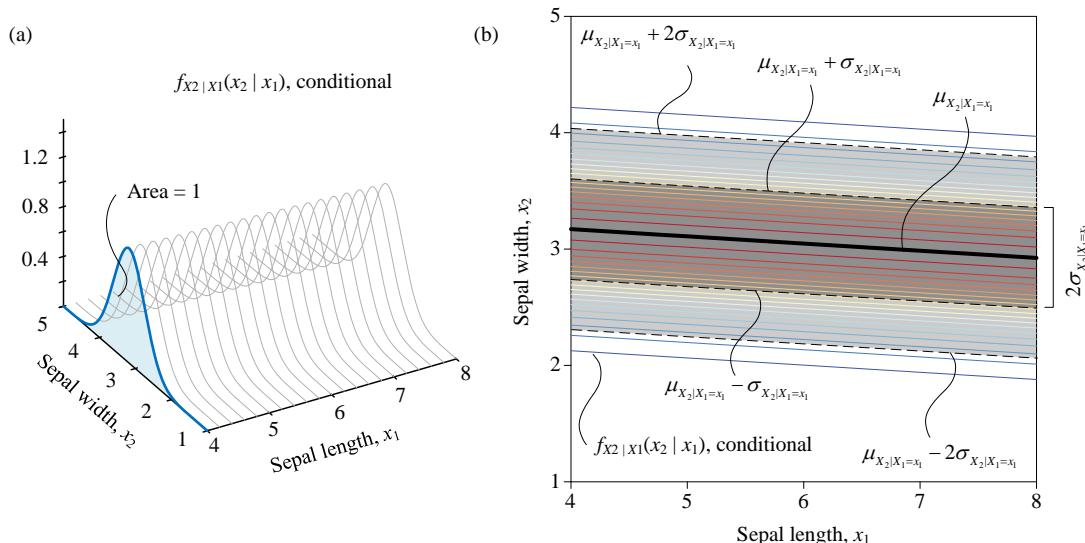
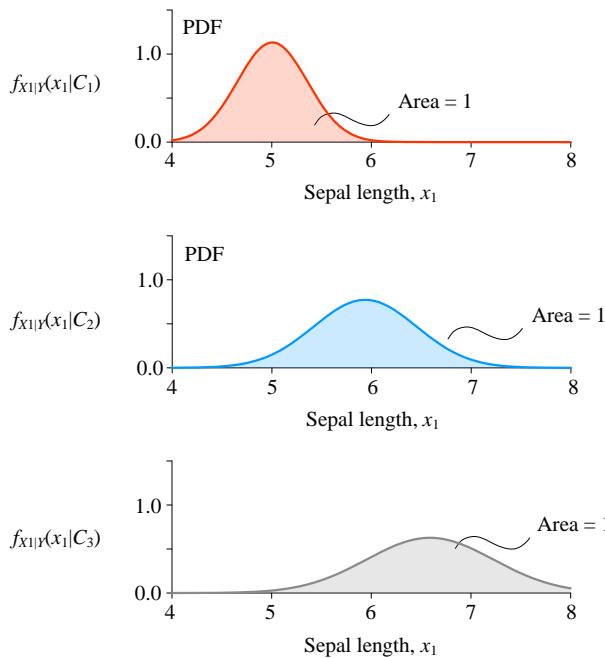


图 12. 条件 PDF $f_{X2|X1}(x_2 | x_1)$ 三维等高线和平面等高线，不考虑分类

以鸢尾花为例，考虑标签

换个条件来看，如图 13 所示，给定鸢尾花分类条件，假设花萼长度服从高斯分布。请大家自行计算给定鸢尾花分类为条件，花萼长度的条件期望 $E(X_1 | Y = C_k)$ 、条件方差 $\text{var}(X_1 | Y = C_k)$ 。

图 13. 给定鸢尾花标签 Y , 花萼长度的 PDF, 连续随机变量

12.3 给定 Y 条件下, X 的条件概率: 以二元高斯为例

如果 (X, Y) 服从二元高斯分布, 给定 $Y=y$ 条件下, X 的条件 PDF $f_{X|Y}(x|y)$ 解析式为:

$$f_{X|Y}(x|y) = \frac{1}{\sigma_X \sqrt{1-\rho_{X,Y}^2} \sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x - (\mu_X + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y} (y - \mu_Y))}{\sigma_X \sqrt{1-\rho_{X,Y}^2}} \right)^2 \right\} \quad (14)$$

图 14 所示为 $f_{X|Y}(x|y)$ 网格线。给定 $Y=y$ 的条件下, 条件 PDF $f_{X|Y}(x|y)$ 投影到 xz 平面上得到图 15。

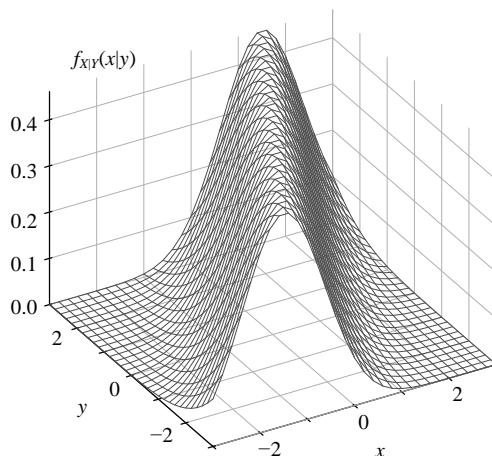
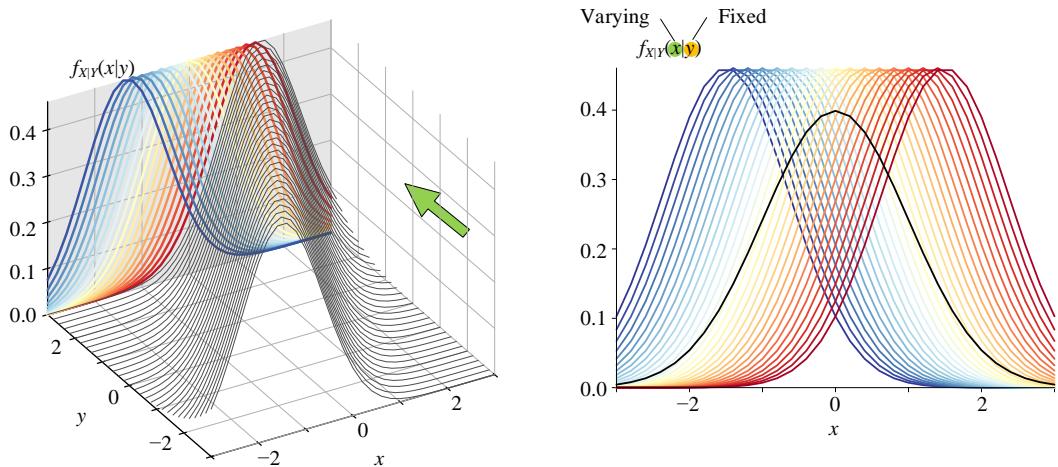
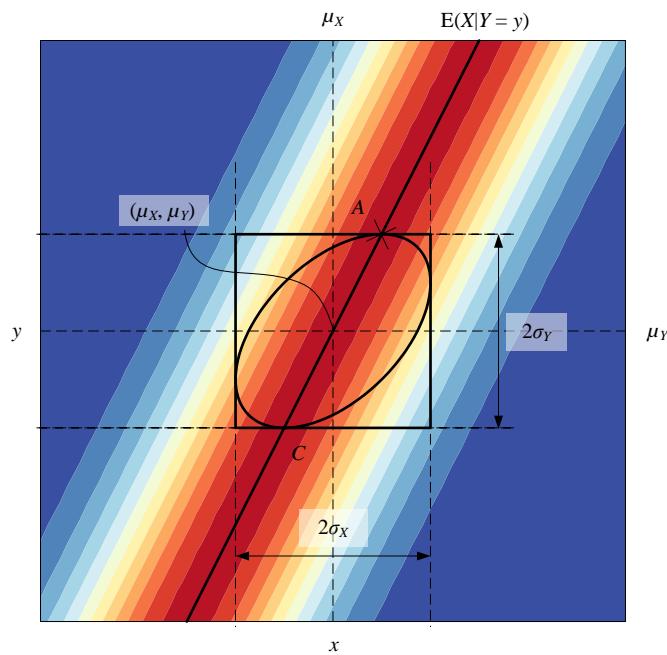


图 14. $f_{X|Y}(x|y)$ 曲面网格线图 15. $f_{X|Y}(x|y)$ 曲面在 xz 平面上投影

条件期望 $E(X|Y=y)$

图 16 所示为 $f_{X|Y}(x|y)$ 的平面等高线。图中的等高线都平行于条件期望 $E(X|Y=y)$ (黑色斜线)，具体解析式为：

$$E(X|Y=y) = \mu_X + \rho_{X,Y} \frac{\sigma_X}{\sigma_Y} (y - \mu_Y) \quad (15)$$

图 16. $f_{X|Y}(x|y)$ 平面等高线

条件方差 $\text{var}(X|Y=y)$

给定 $Y=y$ 条件下， Y 的条件**方差** $\text{var}(X|Y=y)$ 解析式为：

$$\text{var}(X|Y=y) = (1 - \rho_{X,Y}^2) \sigma_X^2 \quad (16)$$

给定 $Y=y$ 条件下， Y 的条件**标准差** $\sigma_{X|Y=y}$ 解析式也是定值：

$$\text{std}(X|Y=y) = \sqrt{(1 - \rho_{X,Y}^2)} \cdot \sigma_X \quad (17)$$

图 17 所示为条件**标准差** $\sigma_{X|Y}$ 的几何含义。

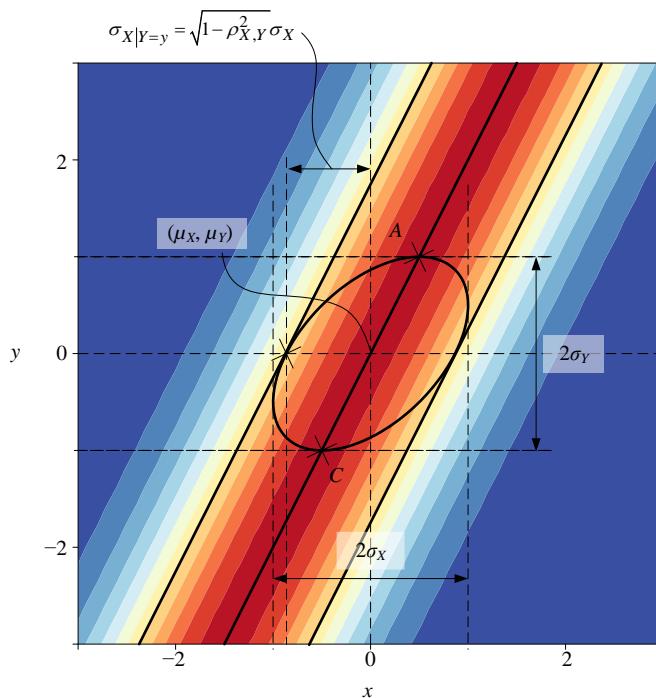


图 17. 条件**标准差** $\sigma_{X|Y}$ 的几何含义

以鸢尾花为例：条件期望 $E(X_1 | X_2=x_2)$ 、条件方差 $\text{var}(X_1 | X_2=x_2)$

以鸢尾花花萼长度 (X_1)、花萼宽度 (X_2) 数据为例，假设 (X_1, X_2) 服从二元高斯分布。给定 $X_2=x_2$ 条件下， X_1 的条件**期望** $E(X_1 | X_2=x_2)$ 解析式为：

$$\begin{aligned}
 E(X_1 | X_2 = x_2) &= \mu_1 + \rho_{1,2} \frac{\sigma_1}{\sigma_2} (x_2 - \mu_2) \\
 &= 5.843 - 0.117 \times \frac{0.825}{0.434} (x_2 - 3.057) \\
 &= -0.222x_2 + 6.523
 \end{aligned} \tag{18}$$

条件方差 $\text{var}(X_1 | X_2 = x_2)$ 解析式为：

$$\text{var}(X_1 | X_2 = x_2) = (1 - \rho_{1,2}^2) \sigma_1^2 \approx 0.671 \tag{19}$$

条件标准差 $\sigma_{X_1 | X_2 = x_2}$ 解析式为定值：

$$\sigma_{X_1 | X_2 = x_2} = \sqrt{1 - \rho_{1,2}^2} \sigma_1 \approx 0.819 \tag{20}$$

类似地，如图 18 所示不管 x_2 怎么变，这个条件标准差为定值。

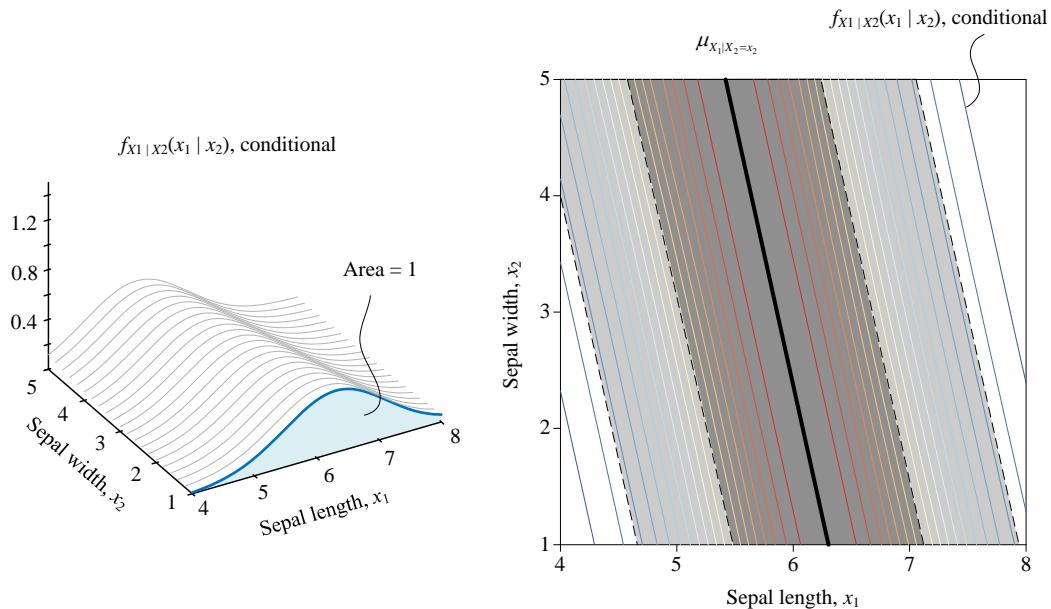
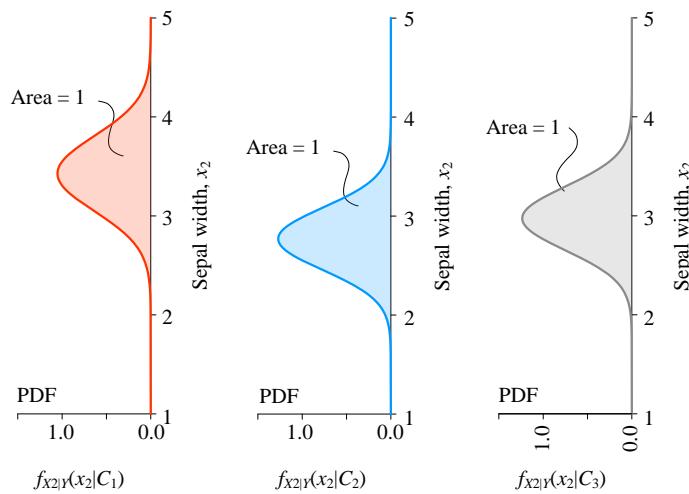


图 18. $f_{X1|X2}(x_1 | x_2)$ 条件 PDF 密度三维等高线和平面等高线，不考虑分类

以鸢尾花为例，考虑标签

换个条件来看，如图 19 所示，给定鸢尾花分类条件，假设花萼宽度服从高斯分布。请大家自行计算给定鸢尾花分类为条件，花萼宽度的条件期望 $E(X_2 | Y = C_k)$ 、条件方差 $\text{var}(X_2 | Y = C_k)$ 。

图 19. 给定鸢尾花标签 Y , 花萼宽度的 PDF 曲线, 连续随机变量

12.4 多元正态条件分布：引入矩阵运算

本节利用矩阵运算讨论多元正态条件分布。

多元高斯分布

如果随机变量向量 χ 和 γ 服从多维高斯分布：

$$\begin{bmatrix} \chi \\ \gamma \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_\chi \\ \mu_\gamma \end{bmatrix}, \begin{bmatrix} \Sigma_{\chi\chi} & \Sigma_{\chi\gamma} \\ \Sigma_{\gamma\chi} & \Sigma_{\gamma\gamma} \end{bmatrix}\right) \quad (21)$$

其中, χ 为随机变量 X_i 构成的列向量, γ 为随机变量 Y_j 构成的列向量:

$$\chi = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix}, \quad \gamma = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_M \end{bmatrix} \quad (22)$$

图 20 所示为多元高斯分布的均值向量、协方差矩阵形状。

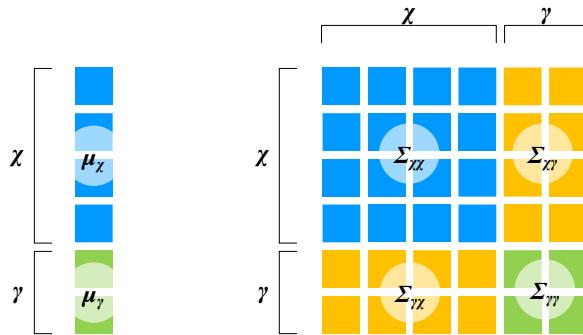


图 20. 均值向量、协方差矩阵形状

互协方差矩阵

注意， $\Sigma_{\gamma\gamma}$ 的转置为 $\Sigma_{\chi\gamma}$ ：

$$(\Sigma_{\gamma\gamma})^T = \Sigma_{\chi\gamma} \quad (23)$$

$\Sigma_{\chi\gamma}$ 也叫互协方差矩阵 (cross-covariance matrix)，这是下一章要讨论的内容之一。

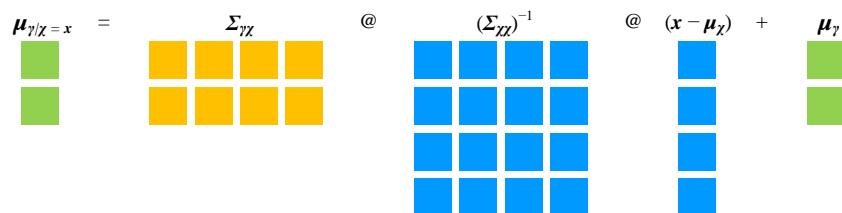
给定 $\chi=x$ 的条件

给定 $\chi=x$ 的条件下， γ 服从如下多维高斯分布：

$$\{\gamma|\chi=x\} \sim N\left(\underbrace{\Sigma_{\gamma\chi}\Sigma_{\chi\chi}^{-1}(x - \mu_\chi)}_{\text{Expectation}}, \underbrace{\Sigma_{\gamma\gamma} - \Sigma_{\gamma\chi}\Sigma_{\chi\chi}^{-1}\Sigma_{\chi\gamma}}_{\text{Covariance matrix}}\right) \quad (24)$$

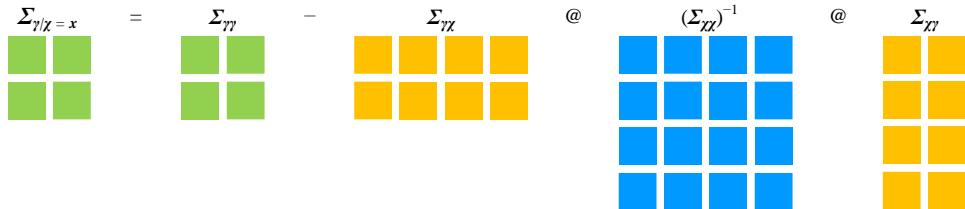
也就是说，如图 21 所示，给定 $\chi=x$ 的条件下 γ 的条件期望为：

$$E(\gamma|\chi=x) = \mu_{\gamma|x=x} = \Sigma_{\gamma\chi}\Sigma_{\chi\chi}^{-1}(x - \mu_\chi) + \mu_\gamma \quad (25)$$

图 21. 给定 $\chi=x$ 的条件下 γ 的期望值的矩阵运算

如图 22 所示，给定 $\chi=x$ 的条件下 γ 的方差为：

$$\boldsymbol{\Sigma}_{\gamma|\chi=x} = \boldsymbol{\Sigma}_{\gamma\gamma} - \boldsymbol{\Sigma}_{\gamma\chi} \boldsymbol{\Sigma}_{\chi\chi}^{-1} \boldsymbol{\Sigma}_{\chi\gamma} \quad (26)$$

图 22. 给定 $\chi=x$ 的条件下 γ 的方差的矩阵运算

给定 $\gamma=y$ 的条件

同理，给定 $\gamma=y$ 的条件下 χ 服从如下多维高斯分布：

$$\{\chi|\gamma=y\} \sim N\left(\underbrace{\boldsymbol{\Sigma}_{\chi\gamma}\boldsymbol{\Sigma}_{\gamma\gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu}_\gamma)+\boldsymbol{\mu}_\chi}_{\text{Expectation}}, \underbrace{\boldsymbol{\Sigma}_{\chi\chi}-\boldsymbol{\Sigma}_{\chi\gamma}\boldsymbol{\Sigma}_{\gamma\gamma}^{-1}\boldsymbol{\Sigma}_{\gamma\chi}}_{\text{Covariance matrix}}\right) \quad (27)$$

即给定 $\gamma=y$ 的条件下 χ 的期望值为：

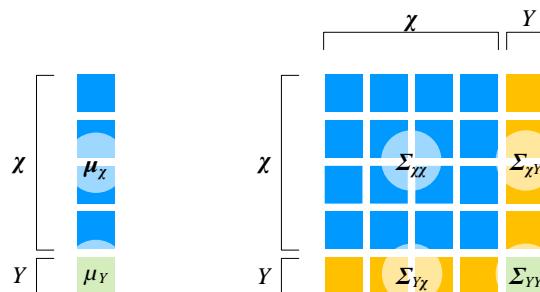
$$\boldsymbol{\mu}_{\chi|\gamma=y} = \boldsymbol{\Sigma}_{\chi\gamma}\boldsymbol{\Sigma}_{\gamma\gamma}^{-1}(\mathbf{y}-\boldsymbol{\mu}_\gamma)+\boldsymbol{\mu}_\chi \quad (28)$$

给定 $\gamma=y$ 的条件下 χ 的方差为：

$$\boldsymbol{\Sigma}_{\chi|\gamma=y} = \boldsymbol{\Sigma}_{\chi\chi} - \boldsymbol{\Sigma}_{\chi\gamma}\boldsymbol{\Sigma}_{\gamma\gamma}^{-1}\boldsymbol{\Sigma}_{\gamma\chi} \quad (29)$$

单一因变量

特别地， γ 只有一个随机变量 Y 时，这对应线性回归中有多个自变量，只有一个因变量，如图 23 所示。

图 23. 均值向量、协方差矩阵形状， γ 只有一个随机变量

这种情况下，给定 $\chi = x$ 条件下 Y 的条件期望为：

$$\mu_{Y|\chi=x} = \Sigma_{Y\chi} \Sigma_{\chi\chi}^{-1} (x - \mu_\chi) + \mu_Y \quad (30)$$

(30) 对应多元线性回归。图 24 对应矩阵运算示意图。

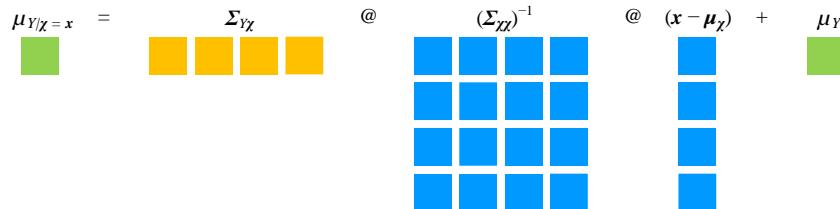


图 24. 给定 $\chi = x$ 条件下 Y 的条件期望

多元线性回归

不考虑常数项系数，如果是行向量表达的话，多元线性回归的系数 b 为：

$$b = [b_1 \ b_2 \ \cdots \ b_D] = \Sigma_{Y\chi} \Sigma_{\chi\chi}^{-1} \quad (31)$$

图 25 所示为 b 的矩阵运算。

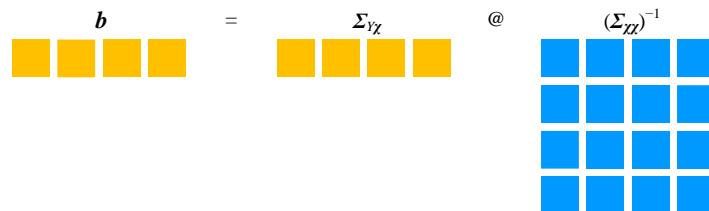


图 25. 计算多元回归的系数 b

常数项 b_0 为：

$$b_0 = -\Sigma_{Y\chi} \Sigma_{\chi\chi}^{-1} \mu_\chi + \mu_Y \quad (32)$$

简单线性回归

更特殊地，当 χ 和 γ 都只有一个随机变量时，即单一自变量 X 、单一因变量 Y ：

$$\mu_{Y|X=x} = \text{cov}(X, Y) (\sigma_x^2)^{-1} (x - \mu_x) + \mu_Y = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_x) + \mu_Y \quad (33)$$

这和本书之前的 (8) 完全一致。本书第 24 章将接续讨论这一话题。

以鸢尾花为例

图 26 所示为鸢尾花数据的质心向量和协方差矩阵热图。我们用花瓣长度、花瓣宽度、萼片长度为多元线性回归的多变量，用花瓣宽度为因变量。图 26 所示向量和协方差矩阵也据此分块。

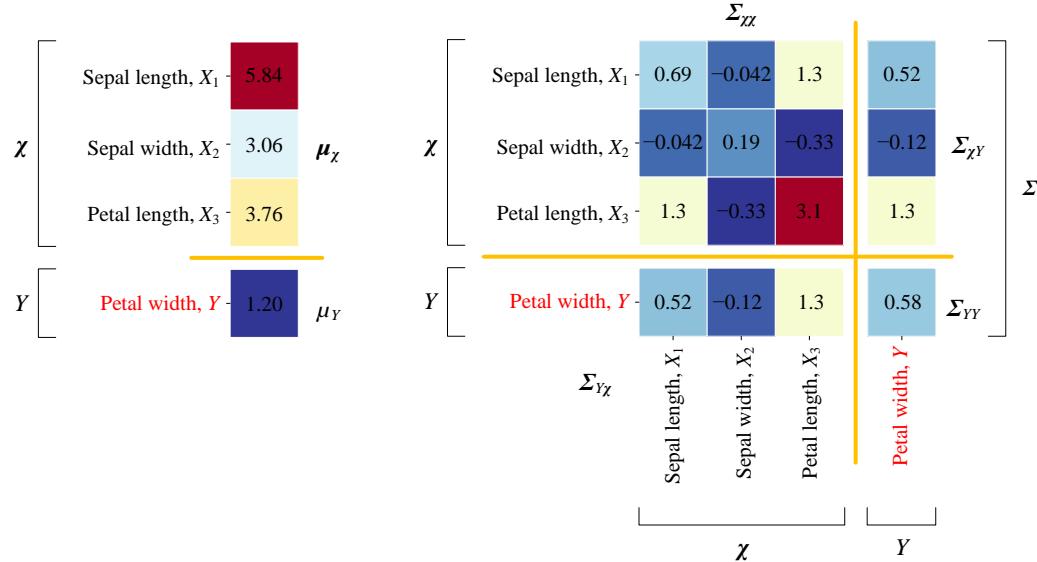


图 26. 质心向量、协方差矩阵热图

利用 (31)，我们可以计算得到回归系数：

$$\begin{aligned}
 \mathbf{b} &= [b_1 \ b_2 \ b_3] = \Sigma_{YX} \Sigma_{XX}^{-1} \\
 &= [0.516 \ -0.122 \ 1.296] \begin{bmatrix} 0.686 & -0.042 & 1.274 \\ -0.042 & 0.190 & -0.330 \\ 1.274 & -0.330 & 3.116 \end{bmatrix}^{-1} \\
 &= [-0.207 \ 0.223 \ 0.524]
 \end{aligned} \tag{34}$$

图 27 所示为上式运算的热图。

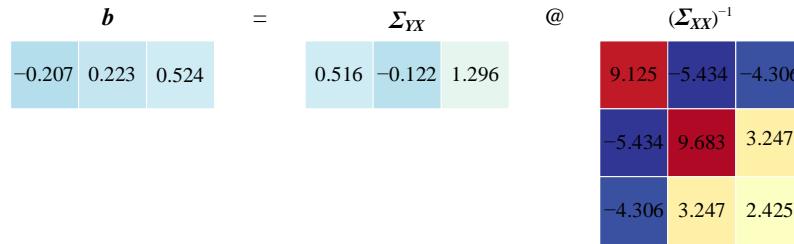


图 27. 矩阵计算系数向量 \mathbf{b}

利用(32)，计算得到多元线性回归的常数项：

$$b_0 = 1.199 - [-0.207 \quad 0.223 \quad 0.524] \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \end{bmatrix} = -0.24 \quad (35)$$

图 28 所示为对应计算。

b_0	=	μ_Y	-	b	@	μ_X	
-0.240		1.199		-0.207	0.223	0.524	
							5.843
							3.057
							3.758

图 28. 矩阵计算常数 b_0

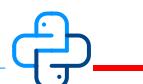
从而得到多元线性回归的解析式：

$$\begin{aligned} y &= [0.516 \quad -0.122 \quad 1.296] \begin{bmatrix} 0.686 & -0.042 & 1.274 \\ -0.042 & 0.190 & -0.330 \\ 1.274 & -0.330 & 3.116 \end{bmatrix}^{-1} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \end{bmatrix} + 1.199 \quad (36) \\ &= -0.207x_1 + 0.223x_2 + 0.524x_3 - 0.240 \end{aligned}$$

这个式子相当于用花萼长度、花萼宽度、花瓣长度作为变量估算花萼宽度。

有必要强调一下，线性回归并不一定表示因果关系。尽管线性回归可以用来探索变量之间的关系，但它并不会告诉我们一个变量是否是另一个变量的原因。因为在统计学中，相关性并不等于因果关系。

要确定两个变量之间的因果关系，需要进行实验研究，例如随机对照实验。在这种类型的实验中，研究人员可以控制潜在的影响因素，然后观察自变量对因变量的影响。因此，线性回归可以用于探索变量之间的关系，但要确定因果关系需要进行更深入的研究和分析。



Bk5_Ch12_03.py 绘制本节图像。



简单来说，条件高斯分布是指在已知某些变量的取值情况下，对另外一些变量的概率分布进行建模的一种方法。条件高斯分布在模式识别、机器学习、贝叶斯推断等领域都有广泛的应用。本节中大家看到条件高斯分布给线性回归提供一种全新的解读视角。

13

Dive into Covariance Matrix

协方差矩阵

很多数学科学、机器学习算法的起点



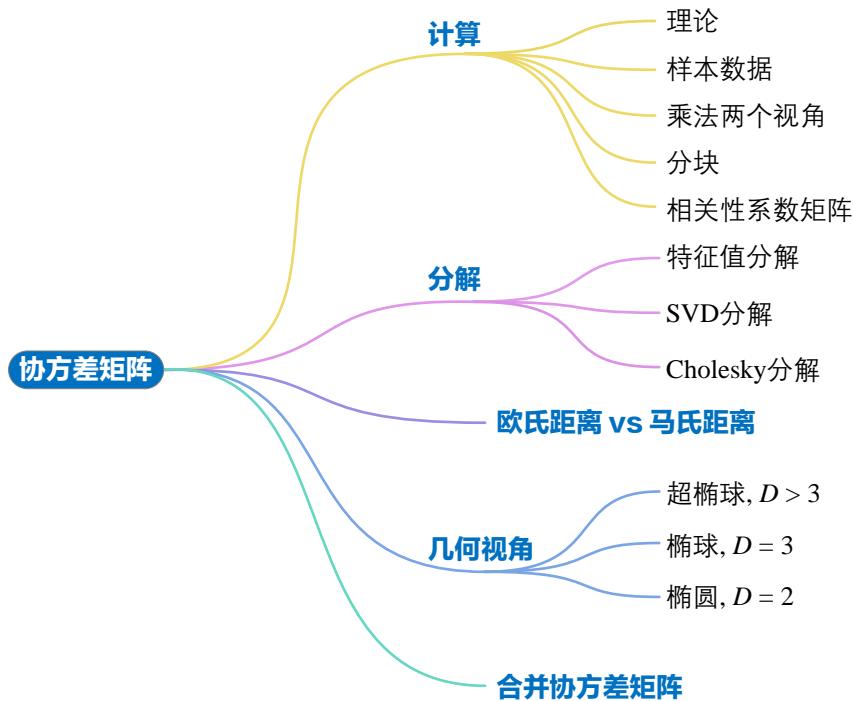
科学的目标是寻求对复杂事实的最简单的解释。我们很容易误以为事实很简单，因为简单是我们追求的目标。每个自然哲学家生活中的指导格言都应该是——寻求简单而不相信它。

The aim of science is to seek the simplest explanations of complex facts. We are apt to fall into the error of thinking that the facts are simple because simplicity is the goal of our quest. The guiding motto in the life of every natural philosopher should be, seek simplicity and distrust it.

——阿尔弗雷德·怀特海 (Alfred Whitehead) | 英国数学家、哲学家 | 1861 ~ 1947



- ◀ `numpy.average()` 计算平均值
- ◀ `numpy.corrcoef()` 计算数据的相关性系数
- ◀ `numpy.cov()` 计算协方差矩阵
- ◀ `numpy.diag()` 如果 A 为方阵, `numpy.diag(A)` 函数提取对角线元素, 以向量形式输入结果; 如果 a 为向量, `numpy.diag(a)` 函数将向量展开成方阵, 方阵对角线元素为 a 向量元素
- ◀ `numpy.linalg.cholesky()` Cholesky 分解
- ◀ `numpy.linalg.eig()` 特征值分解
- ◀ `numpy.linalg.inv()` 矩阵求逆
- ◀ `numpy.linalg.norm()` 计算范数
- ◀ `numpy.linalg.svd()` 奇异值分解
- ◀ `numpy.ones()` 创建全 1 向量或矩阵
- ◀ `numpy.sqrt()` 计算平方根



13.1 计算协方差矩阵：描述数据分布

协方差矩阵囊括多特征数据矩阵重要统计描述，在多元高斯分布中协方差矩阵扮演重要角色。不仅如此，数据科学和机器学习方法中随处可见，比如多元高斯分布、随机数生成器、OLS 线性回归、主成分分析、正交回归、高斯过程、高斯朴素贝叶斯、高斯判别分析、高斯混合模型等。因此，我们有必要拿一章专门讨论协方差矩阵。

本系列丛书介绍的很多数学概念在协方差矩阵处达到完美融合，比如解析几何中的椭圆，概率统计中的高斯分布，线性代数中的线性变换、Cholesky 分解、特征值分解、正定性等。因此，本章也可以视作是对《矩阵力量》中重要的线性代数工具的梳理和应用。

形状

一般而言，协方差矩阵可视作方差和协方差两部分组成，方差是协方差矩阵对角线上的元素，协方差是协方差矩阵非对角线上的元素：

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D,1} & \sigma_{D,2} & \cdots & \sigma_{D,D} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,D}\sigma_1\sigma_D \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,D}\sigma_2\sigma_D \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D}\sigma_1\sigma_D & \rho_{2,D}\sigma_2\sigma_D & \cdots & \sigma_D^2 \end{bmatrix} \quad (1)$$

方差描述了某个特征上数据的离散度，而协方差则蕴含成对特征之间的相关性。

显而易见，协方差矩阵为对称矩阵：

$$\Sigma = \Sigma^T \quad (2)$$

理论

定义随机变量的列向量 χ ：

$$\chi = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix} \quad (3)$$

χ 的协方差矩阵可以通过下式计算得到：

$$\begin{aligned} \text{var}(\chi) &= \text{cov}(\chi, \chi) = E[(\chi - E(\chi))(\chi - E(\chi))^T] \\ &= E(\chi \chi^T) - E(\chi)E(\chi)^T \end{aligned} \quad (4)$$

⚠ 注意，为了方便表达，上式中列向量 χ 期望值向量 $E(\chi)$ 也是列向量。 $E(\chi\chi^T)$ 和 $E(\chi)E(\chi)^T$ 的结果都是 $D \times D$ 方阵。

上式类似我们在本书第 4 章提到的计算方差和协方差的技巧，请大家类比：

$$\begin{aligned} \text{var}(X) &= \underbrace{E(X^2)}_{\text{Expectation of } X^2} - \underbrace{E(X)^2}_{\text{Square of } E(X)} \\ \text{cov}(X_1, X_2) &= E(X_1 X_2) - E(X_1)E(X_2) \end{aligned} \quad (5)$$

样本数据

实践中，我们更常用的是样本数据的协方差矩阵。对于形状为 $n \times D$ 的样本数据矩阵 X ， X 的协方差矩阵 Σ 可以通过下式计算得到：

$$\Sigma = \frac{\left(\underbrace{X - E(X)}_{\text{Centered}} \right)^T \left(\underbrace{X - E(X)}_{\text{Centered}} \right)}{n-1} = \frac{X_c^T X_c}{n-1} \quad (6)$$

其中， $E(X)$ 为数据 X 质心，是行向量；利用广播原则， $X - E(X)$ 得到去均值数据矩阵 X_c 。

⚠ 注意，式中分母为 $n - 1$ 。

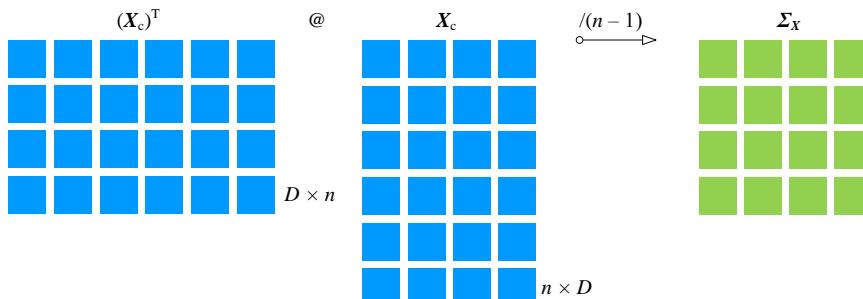


图 1. 计算 X 样本数据协方差矩阵 Σ_X

(6) 可以写成：

$$\Sigma = \frac{(X - I E(X))^T (X - I E(X))}{n-1} \quad (7)$$

(7) 展开得到：

$$\begin{aligned}
 \Sigma &= \frac{(\mathbf{X}^T - \mathbb{E}(\mathbf{X})^T \mathbf{I}^T)(\mathbf{X} - \mathbf{I}\mathbb{E}(\mathbf{X}))}{n-1} \\
 &= \frac{\mathbf{X}^T \mathbf{X} - \mathbb{E}(\mathbf{X})^T \mathbf{I}^T \mathbf{X} - \mathbf{X}^T \mathbf{I} \mathbb{E}(\mathbf{X}) + \mathbb{E}(\mathbf{X})^T \mathbf{I}^T \mathbf{I} \mathbb{E}(\mathbf{X})}{n-1} \\
 &= \frac{\mathbf{X}^T \mathbf{X}}{n-1} - \frac{n}{n-1} \mathbb{E}(\mathbf{X})^T \mathbb{E}(\mathbf{X})
 \end{aligned} \tag{8}$$

Gram matrix

观察(8)，相信大家已经看到**格拉姆矩阵**(Gram matrix)。也就是说，协方差矩阵可以视作一种特殊的格拉姆矩阵。

此外，如果 n 足够大，可以用 n 替换 $n-1$ ，影响微乎其微。

把数据矩阵 \mathbf{X} 展开成一组列向量 $[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D]$ ， $\mathbb{E}(\mathbf{X})$ 写成 $[\mu_1, \mu_2, \dots, \mu_D]$ ，(6) 可以整理为：

$$\begin{aligned}
 \Sigma &= \frac{(\mathbf{X} - \mathbb{E}(\mathbf{X}))^T (\mathbf{X} - \mathbb{E}(\mathbf{X}))}{n-1} \\
 &= \frac{[\mathbf{x}_1 - \mu_1 \quad \mathbf{x}_2 - \mu_2 \quad \cdots \quad \mathbf{x}_D - \mu_D]^T [\mathbf{x}_1 - \mu_1 \quad \mathbf{x}_2 - \mu_2 \quad \cdots \quad \mathbf{x}_D - \mu_D]}{n-1} \\
 &= \frac{1}{n-1} \begin{bmatrix} (\mathbf{x}_1 - \mu_1)^T (\mathbf{x}_1 - \mu_1) & (\mathbf{x}_1 - \mu_1)^T (\mathbf{x}_2 - \mu_2) & \cdots & (\mathbf{x}_1 - \mu_1)^T (\mathbf{x}_D - \mu_D) \\ (\mathbf{x}_2 - \mu_2)^T (\mathbf{x}_1 - \mu_1) & (\mathbf{x}_2 - \mu_2)^T (\mathbf{x}_2 - \mu_2) & \cdots & (\mathbf{x}_2 - \mu_2)^T (\mathbf{x}_D - \mu_D) \\ \vdots & \vdots & \ddots & \vdots \\ (\mathbf{x}_D - \mu_D)^T (\mathbf{x}_1 - \mu_1) & (\mathbf{x}_D - \mu_D)^T (\mathbf{x}_2 - \mu_2) & \cdots & (\mathbf{x}_D - \mu_D)^T (\mathbf{x}_D - \mu_D) \end{bmatrix}
 \end{aligned} \tag{9}$$

图 2 (a) 所示为鸢尾花四特征数据协方差矩阵 Σ 。

上一章讲解多元高斯分布时，讲过其概率密度函数 PDF 解析式中用到协方差矩阵的逆。而协方差矩阵的逆矩阵有自己的名字——**集中矩阵**(concentration matrix)。图 2 (b) 所示为协方差矩阵的逆 Σ^{-1} 。

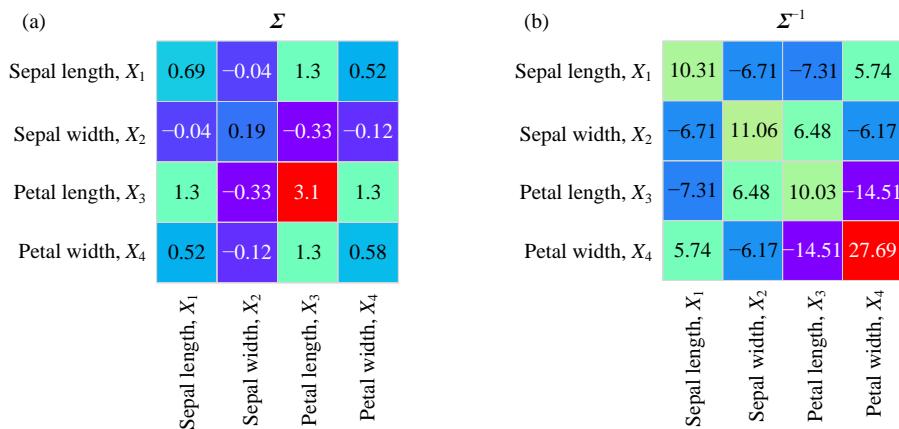


图 2. 鸢尾花四特征协方差矩阵、逆矩阵热图

四种椭圆

本书中常用椭圆代表协方差矩阵。 χ 若服从多元高斯分布， $\chi \sim (\mu, \Sigma)$ 。如图 3 所示，当协方差矩阵形态不同时，对应的椭圆有四种类型。

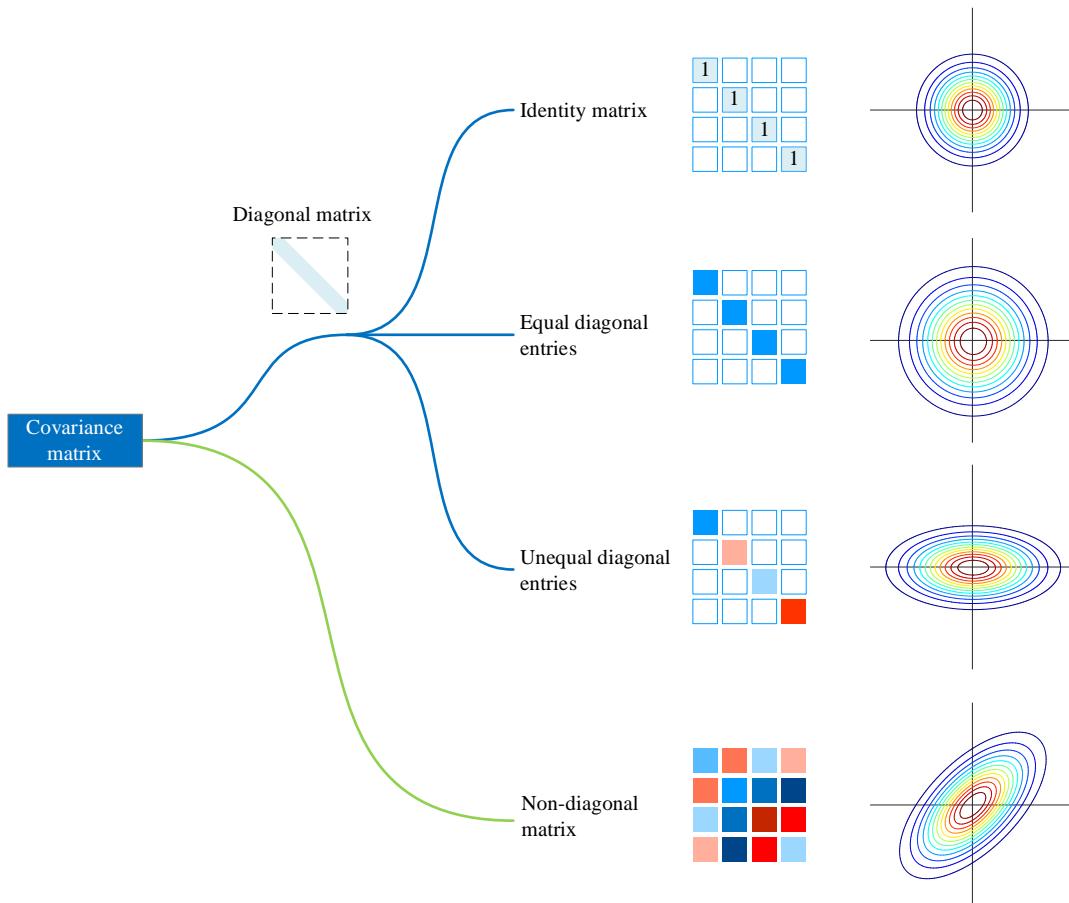


图 3. 协方差矩阵的形态影响高斯密度函数形状

当协方差矩阵为**单位矩阵** (identity matrix) 时，即 $\Sigma = I$ ，随机变量为 IID，每一个随机变量服从标准正态分布；因此，这种情况，我们用正圆代表其概率密度函数。准确来说是，概率密度函数对应的几何形状是多维空间的正球体。

独立同分布 (Independent and identically distributed, IID) 是指一组随机变量中每个变量的概率分布都相同，且这些随机变量互相独立。

类似的，当协方差矩阵为 $\Sigma = kI$ ，这种情况对应的概率密度函数也是正圆， k 相当于缩放系数。

当 Σ 为对角阵，对角线元素不同：

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & 0 & \cdots & 0 \\ 0 & \sigma_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{D,D} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D^2 \end{bmatrix} \quad (10)$$

这种情况，对应的概率密度函数形状为正椭圆。多元高斯分布的密度函数可以写成边际概率密度函数的累乘：

$$f_X(\mathbf{x}) = \prod_{j=1}^D \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{1}{2}\left(\frac{x_j - \mu_j}{\sigma_j}\right)^2\right) \quad (11)$$

Σ 不定时，高斯分布 PDF 形状为旋转椭圆。

本章最后将深入探讨协方差的几何视角。

给定标签为条件

当然，在计算协方差时，我们也可以考虑到数据标签。图 4 所示为三个不同标签数据各自的协方差矩阵 $\Sigma_1, \Sigma_2, \Sigma_3$ 热图。

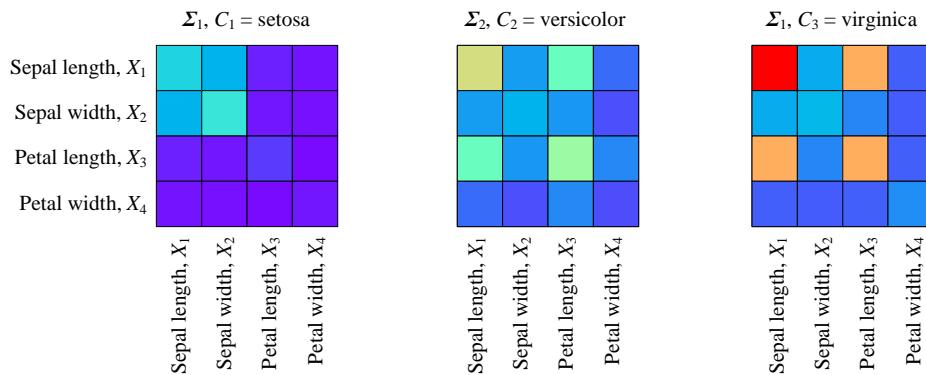


图 4. 协方差矩阵热图，考虑分类

质心位于原点

特别地，当所有均值都是 0 时， $[\mu_1, \mu_2, \dots, \mu_D]^T = [0, 0, \dots, 0]^T$ ，也就是说数据质心位于原点，并将 X 写成列向量，(9) 可以写成：

$$\Sigma = \frac{\mathbf{X}^T \mathbf{X}}{n-1} = \frac{\mathbf{G}}{n-1} = \frac{1}{n-1} \begin{bmatrix} \mathbf{x}_1^T \mathbf{x}_1 & \mathbf{x}_1^T \mathbf{x}_2 & \cdots & \mathbf{x}_1^T \mathbf{x}_D \\ \mathbf{x}_2^T \mathbf{x}_1 & \mathbf{x}_2^T \mathbf{x}_2 & \cdots & \mathbf{x}_2^T \mathbf{x}_D \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_D^T \mathbf{x}_1 & \mathbf{x}_D^T \mathbf{x}_2 & \cdots & \mathbf{x}_D^T \mathbf{x}_D \end{bmatrix} \quad (12)$$

用向量内积运算，(12) 可以写成：

$$\Sigma = \frac{1}{n-1} \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x}_1 \rangle & \langle \mathbf{x}_1, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_1, \mathbf{x}_D \rangle \\ \langle \mathbf{x}_2, \mathbf{x}_1 \rangle & \langle \mathbf{x}_2, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_2, \mathbf{x}_D \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{x}_D, \mathbf{x}_1 \rangle & \langle \mathbf{x}_D, \mathbf{x}_2 \rangle & \cdots & \langle \mathbf{x}_D, \mathbf{x}_D \rangle \end{bmatrix} \quad (13)$$

上式是矩阵乘法的第一视角。

同样，当数据质心位于原点时，将 \mathbf{X} 写成行向量，(9) 可以写成：

$$\begin{aligned} \Sigma &= \frac{\mathbf{X}^T \mathbf{X}}{n-1} = \frac{1}{n-1} \begin{bmatrix} \mathbf{x}^{(1)T} & \mathbf{x}^{(2)T} & \cdots & \mathbf{x}^{(n)T} \end{bmatrix} \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix} \\ &= \frac{1}{n-1} (\mathbf{x}^{(1)T} \mathbf{x}^{(1)} + \mathbf{x}^{(2)T} \mathbf{x}^{(2)} + \cdots + \mathbf{x}^{(n)T} \mathbf{x}^{(n)}) = \frac{1}{n-1} \sum_{i=1}^n \mathbf{x}^{(i)T} \mathbf{x}^{(i)} \end{aligned} \quad (14)$$

上式中， $\mathbf{x}^{(i)T} \mathbf{x}^{(i)}$ 的形状为 $D \times D$ 。矩阵乘法写成 n 个形状大小相同矩阵层层叠加，这便是矩阵乘法的第二视角。

协方差矩阵分块

协方差矩阵还可以分块。比如，鸢尾花 4×4 协方差矩阵可以按照如下方式分块：

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} & \sigma_{1,4} \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} & \sigma_{2,4} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_{3,3} & \sigma_{3,4} \\ \sigma_{4,1} & \sigma_{4,2} & \sigma_{4,3} & \sigma_{4,4} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix} & \begin{bmatrix} \sigma_{1,3} & \sigma_{1,4} \\ \sigma_{2,3} & \sigma_{2,4} \end{bmatrix} \\ \begin{bmatrix} \sigma_{3,1} & \sigma_{3,2} \\ \sigma_{4,1} & \sigma_{4,2} \end{bmatrix} & \begin{bmatrix} \sigma_{3,3} & \sigma_{3,4} \\ \sigma_{4,3} & \sigma_{4,4} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \Sigma_{2 \times 2} & \Sigma_{2 \times (4-2)} \\ \Sigma_{(4-2) \times 2} & \Sigma_{(4-2) \times (4-2)} \end{bmatrix} \quad (15)$$

4×4 协方差矩阵 Σ 被分为 4 块。注意，矩阵分块时切割线的交点位于主对角线上。

如图 5 所示， $\Sigma_{2 \times 2}$ 和 $\Sigma_{(4-2) \times (4-2)}$ 都还是协方差矩阵，它俩的主对角线上还是方差。几何视角来看， $\Sigma_{2 \times 2}$ 和 $\Sigma_{(4-2) \times (4-2)}$ 都是旋转椭圆。

而 $\Sigma_{(4-2) \times 2}$ 和 $\Sigma_{2 \times (4-2)}$ 叫互协方差矩阵 (cross-covariance matrix)。

⚠ 注意，互协方差矩阵中一般只含有协方差，没有方差。

$\Sigma_{(4-2) \times 2}$ 和 $\Sigma_{2 \times (4-2)}$ 互为转置矩阵，即 $\Sigma_{(4-2) \times 2} = \Sigma_{2 \times (4-2)}^T$ 。

→ 丛书《数据有道》一册讲解典型相关分析 (Canonical Correlation Analysis) 将会用到互协方差矩阵。

当然，协方差矩阵分块方式有很多，比如图 6。图 6 中 $\Sigma_{3 \times 3}$ 的几何形状为椭球。请大家自行分析图 6。

有关分块矩阵运算，建议大家回顾《矩阵力量》第 6 章。

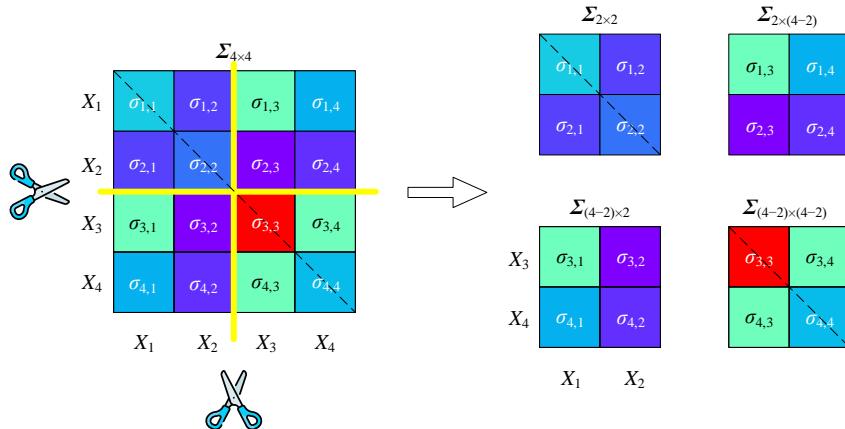


图 5. 协方差矩阵分块

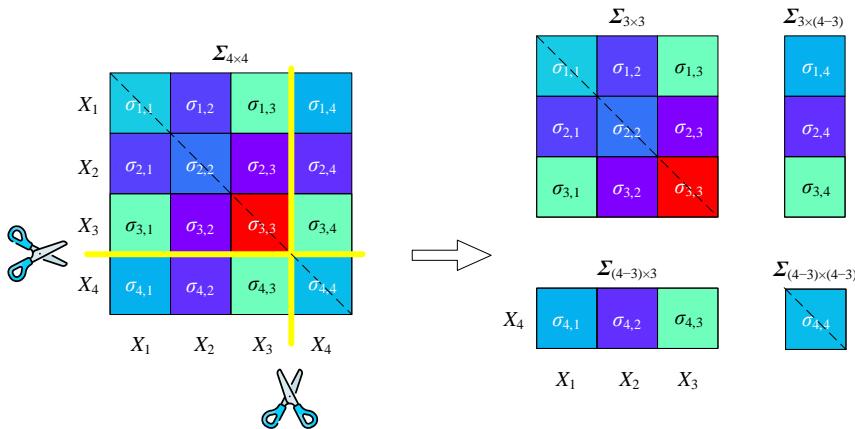


图 6. 协方差矩阵分块，第二种方式

13.2 相关性系数矩阵：描述 Z 分数分布

相关性系数矩阵 P 的定义为：

$$P = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix} \quad (16)$$

图 7 所示为鸢尾花数据相关性系数矩阵 \mathbf{P} 。 \mathbf{P} 的对角线元素均为 1，对角线以外元素为成对相关性系数 $\rho_{i,j}$ 。类似协方差矩阵，相关性系数矩阵 \mathbf{P} 当然也可以分块。

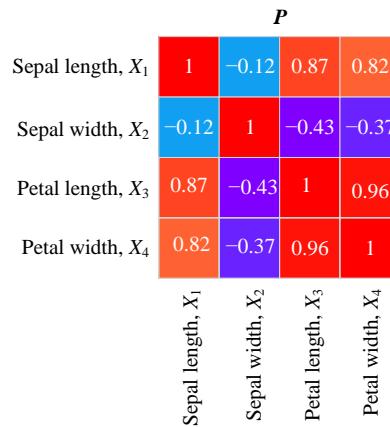


图 7. 鸢尾花数据相关性系数矩阵热图

协方差矩阵 vs 相关性系数矩阵

协方差矩阵 Σ 和相关性系数矩阵 \mathbf{P} 关系如下：

$$\Sigma = \mathbf{D} \mathbf{P} \mathbf{D} = \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix}}_D \underbrace{\begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix}}_{\text{Correlation matrix, } \mathbf{P}} \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix}}_D \quad (17)$$

从几何角度来看，上式中对角方阵 \mathbf{D} 起到的是缩放作用。

图 8 所示为协方差矩阵和相关性矩阵关系热图。

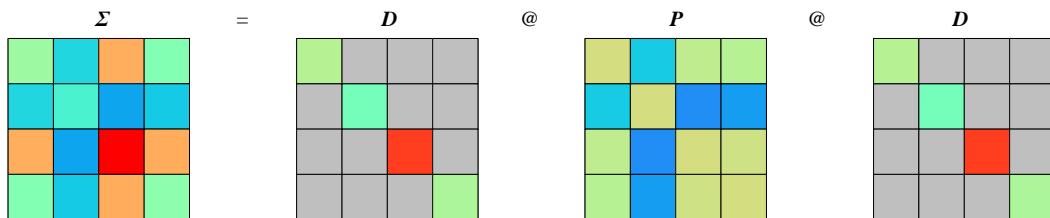


图 8. 协方差矩阵和相关性矩阵关系热图

从 Σ 反求相关性系数矩阵 \mathbf{P}

$$\mathbf{P} = \mathbf{D}^{-1} \Sigma \mathbf{D}^{-1} \quad (18)$$

其中

$$\mathbf{D}^{-1} = \text{diag}\left(\text{diag}(\boldsymbol{\Sigma})\right)^{-1} = \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_D \end{bmatrix} \quad (19)$$

上式中，里层的 `diag()` 提取协方差矩阵的对角线元素（方差），结果为向量。外层的 `diag()` 将向量展成对角方阵。

考虑标签

图 9 为考虑分类标签条件下的协方差矩阵热图，我们管它们叫条件协方差矩阵。

大家是否立刻想到，既然协方差可以用椭圆代表，图 9 中的三个条件协方差矩阵也肯定有它们各自的椭圆！这是本章最后要介绍的内容。

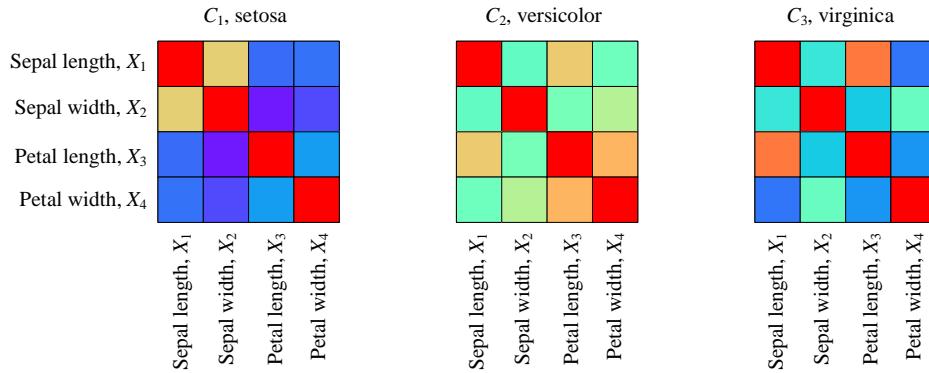


图 9. 相关性系数矩阵热图，考虑分类标签

13.3 特征值分解：找到旋转、缩放

对协方差矩阵 $\boldsymbol{\Sigma}$ 特征值分解：

$$\boldsymbol{\Sigma} = \mathbf{V} \mathbf{A} \mathbf{V}^{-1} \quad (20)$$

其中，特征值矩阵 \mathbf{A} 为对角方阵：

$$\mathbf{A} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} \quad (21)$$

由于 Σ 为对称矩阵，所以对协方差矩阵特征值分解是谱分解：

$$\Sigma = V \Lambda V^T \quad (22)$$

图 10 所示为鸳尾花数据协方差矩阵 Σ 的特征值分解运算热图。

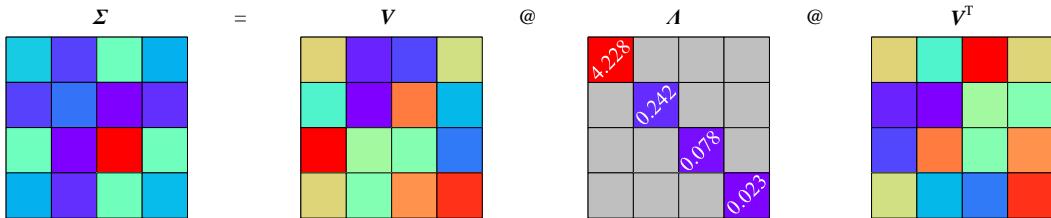


图 10. 协方差矩阵特征值分解

矩阵 V 为正交矩阵：

$$V V^T = I \quad (23)$$

图 11 运算热图对应 (23)。

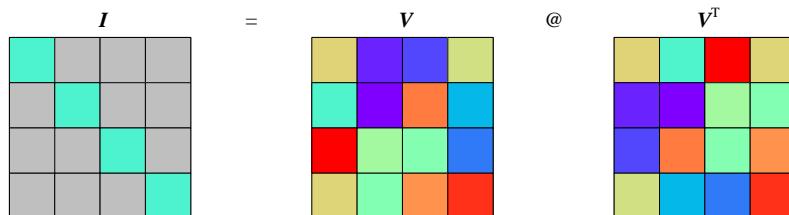


图 11. 矩阵 V 为正交矩阵

谱分解：外积展开

将 (22) 展开来写得到：

$$\begin{aligned} \Sigma &= V \Lambda V^T = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_D] \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix} \\ &= \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \lambda_D \mathbf{v}_D \mathbf{v}_D^T = \sum_{j=1}^D \lambda_j \mathbf{v}_j \mathbf{v}_j^T \end{aligned} \quad (24)$$

这便是《矩阵力量》第 5 章介绍的矩阵乘法第二视角——外积展开，将矩阵乘法展开写成加法。

用向量张量积来写 (24) 得到：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\Sigma = \lambda_1 v_1 \otimes v_1 + \lambda_2 v_2 \otimes v_2 + \cdots + \lambda_D v_D \otimes v_D = \sum_{j=1}^D \lambda_j v_j \otimes v_j \quad (25)$$

注意， v_j 为单位向量，无量纲，即没有单位。

几何角度来看， v_j 仅仅提供投影的方向，而真正提供缩放大小的是特征值 λ_j 。图 12 所示为协方差矩阵谱分解展开热图。虽然 $\lambda_1 v_1 v_1^\top$ 的秩为 1，但是 $\lambda_1 v_1 v_1^\top$ 已经几乎“还原” Σ 。

此外，几何视角来看， $\lambda_1 v_1 v_1^\top$ 代表向量投影，即《矩阵力量》第 10 章中讲过的“二次投影”，建议大家回顾。

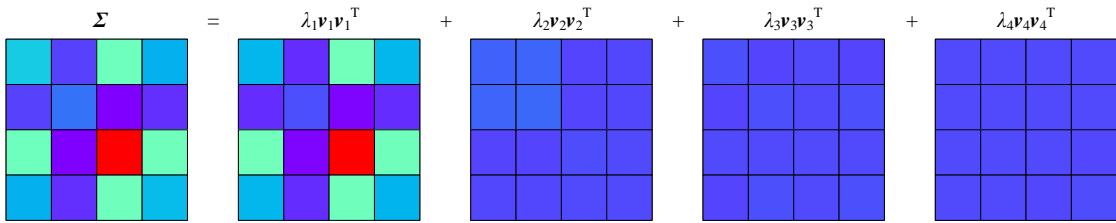


图 12. 协方差矩阵谱分解展开热图

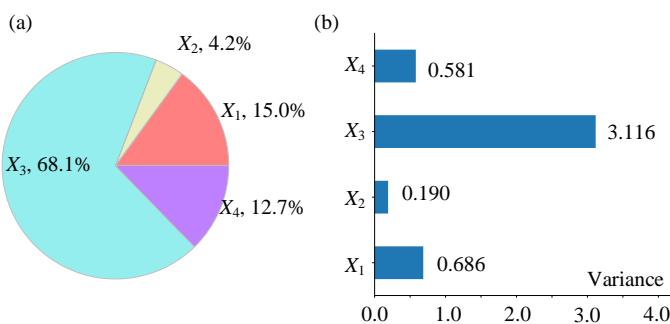
迹

一个值得注意的性质是，协方差矩阵 Σ 的迹——方阵对角线元素之和——等于(21) 特征值之和：

$$\begin{aligned} \text{trace}(\Sigma) &= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_D^2 = \sum_{j=1}^D \sigma_j^2 \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_D = \sum_{j=1}^D \lambda_j \end{aligned} \quad (26)$$

协方差矩阵 Σ 对角线元素之和，相当于所有特征的方差之和，即数据整体的方差。 V 相当于旋转，而旋转操作不改变数据整体方差。本章后文将介绍理解上式的几何视角。

图 13 所示为鸢尾花数据矩阵 X 中每一列数据的方差 σ_j^2 对整体方差 $\sum_{j=1}^D \sigma_j^2$ 的贡献。

图 13. 协方差矩阵 Σ 的主对角线成分，即 X 的方差

投影视角

利用我们已经学过的有关特征值分解的几何视角，中心化数据矩阵 X_c 在 V 投影得到数据 Y ：

$$Y = X_c V = (X - E(X))V \quad (27)$$

求数据矩阵 Y 的协方差矩阵：

$$\begin{aligned} \Sigma_Y &= \frac{Y^T Y}{n-1} = \frac{((X - E(X))V)^T (X - E(X))V}{n-1} \\ &= V^T \frac{(X - E(X))^T (X - E(X))}{n-1} V \\ &= V^T \Sigma V = A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} \end{aligned} \quad (28)$$

观察 (28) 的矩阵 Y 的协方差矩阵，可以发现投影得到的数据列向量相互正交特征值从大到小排列，即 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$ ，矩阵 Y 第一列 y_1 的方差最大。

如图 14 所示，以鸢尾花数据投影结果为例， y_1 的方差对整体方差贡献超过 90%。



这便是主成分分析的思路，本书第 25 章将继续这一话题的探讨。

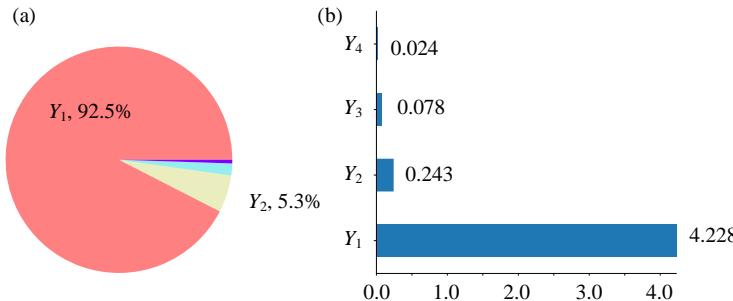


图 14. Σ_Y 的主对角线成分， Y 的方差

协方差的“投影”

举个例子，数据矩阵 X_c 在 v_1 方向投影结果为 y_1 ：

$$y_1 = X_c v_1 \quad (29)$$

由于 X_c 的质心在原点，所以 y_1 的期望值为 0。而 y_1 的方差为：

$$\sigma_{y_1}^2 = \frac{y_1^T y_1}{n-1} = \frac{(X_c v_1)^T X_c v_1}{n-1} = v_1^T \frac{X_c^T X_c}{n-1} v_1 = v_1^T \Sigma v_1 \quad (30)$$

将 (24) 代入上式得到：

$$\Sigma_{y_1} = v_1^T (\lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \dots + \lambda_D v_D v_D^T) v_1 = \lambda_1 \quad (31)$$

上式相当于 Σ 在 v_1 方向上“投影”的结果。

类似地， Σ 在 $[v_1, v_2]$ “投影”的结果为：

$$\begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix} \Sigma [v_1 \quad v_2] = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} \quad (32)$$



本书下一章将深入探讨这一话题。

开平方

用特征值分解结果，可以对协方差矩阵 Σ 开平方：

$$\Sigma = V A^{\frac{1}{2}} A^{\frac{1}{2}} V^T = V A^{\frac{1}{2}} \left(V A^{\frac{1}{2}} \right)^T \quad (33)$$

请大家利用本章代码自行绘制上式热图。

行列式值

协方差矩阵 Σ 的行列式值为其特征值乘积：

$$|\Sigma| = |\Lambda| = \prod_{j=1}^D \lambda_j \quad (34)$$

本章后文会探讨上式的几何内涵。

Σ 行列式值的平方根为：

$$|\Sigma|^{\frac{1}{2}} = |A|^{\frac{1}{2}} = \sqrt{\prod_{j=1}^D \lambda_j} \quad (35)$$

注意，只有在特征值均不为 0 时 $|\Sigma|^{-\frac{1}{2}}$ 才存在，也就是说此时 Σ 为正定。

逆的特征值分解

如果协方差矩阵正定，对协方差矩阵的逆矩阵进行特征值分解，得到：

$$\Sigma^{-1} = (V A V^T)^{-1} = (V^T)^{-1} A^{-1} V^{-1} = V A^{-1} V^T \quad (36)$$

上式利用到对称矩阵特征值分解， $V V^T = I$ 这个性质。

Σ^{-1} 的特征值矩阵为：

$$A^{-1} = \begin{bmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\lambda_D \end{bmatrix} \quad (37)$$

图 15 所示为 Σ^{-1} 的特征值分解运算热图。

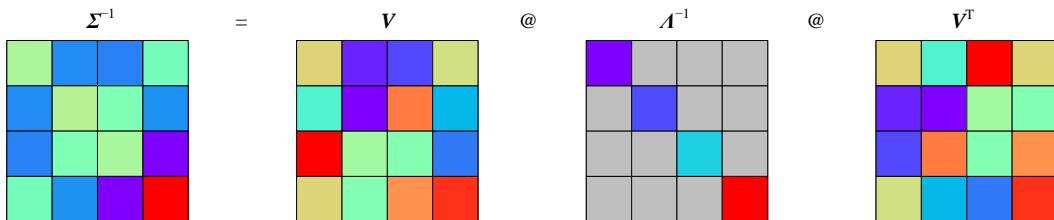


图 15. 协方差矩阵的逆的特征值分解运算热图

相关性系数矩阵的特征值分解

大家肯定能够想到，既然协方差矩阵可以特征值分解，相关性系数矩阵当然也可以进行特征值分解！图 16 所示为相关性系数矩阵的特征值分解，也是谱分解。

对 X 的每一列求 Z 分数得到 Z_X ，相关性系数矩阵是 Z_X 的协方差矩阵。也就是说，如图 16 所示， Z_X 的整体方差为 4。比较图 10 和图 16，容易发现两个正交矩阵不同。

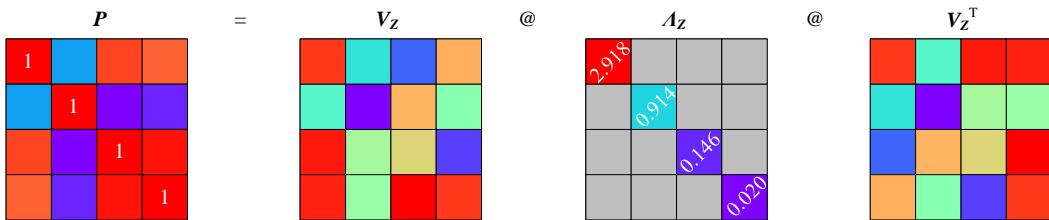


图 16. 相关系数矩阵的特征值分解

13.4 SVD 分解：分解数据矩阵

《矩阵力量》一册反复提过特征值分解 EVD 和奇异值分解 SVD 的关系。本节探讨对中心化 X_c 矩阵 SVD 分解结果和本章前文介绍的特征值分解结果之间的关系。

回顾 SVD 分解

如图 17 所示，对中心化数据矩阵 X_c 进行经济型 SVD 分解得到：

$$X_c = USV^T \quad (38)$$

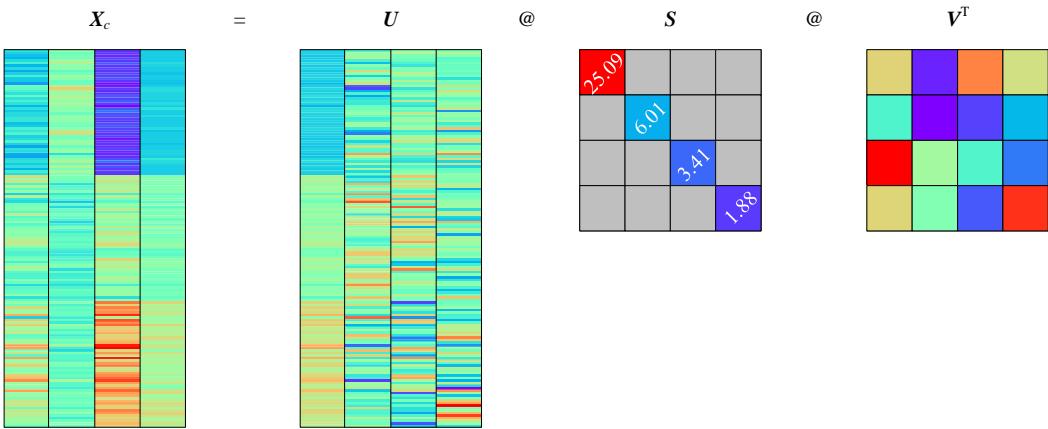
经济型 SVD 分解中， U 的形状和 X_c 完全相同，都是 $n \times D$ 。 U 的列向量两两正交，即满足 $U^T U = I_{D \times D}$ ，但是不满足 $U U^T = I_{n \times n}$ 。

完全型 SVD 分解中， U 的形状为 $n \times n$ 。 U 为正交矩阵，满足 $U^T U = U U^T = I_{n \times n}$ 。

经济型 SVD 分解中， S 为对角方阵，对角元素为奇异值 s_i 。

经济型 SVD 分解中， V 的形状为 $D \times D$ 。 V 为正交矩阵，满足 $V^T V = V V^T = I_{D \times D}$ 。 V 为规范正交基。

注意，本书后文为了区分不同规范正交基，会把 (38) 中的 V 写成 V_c 。

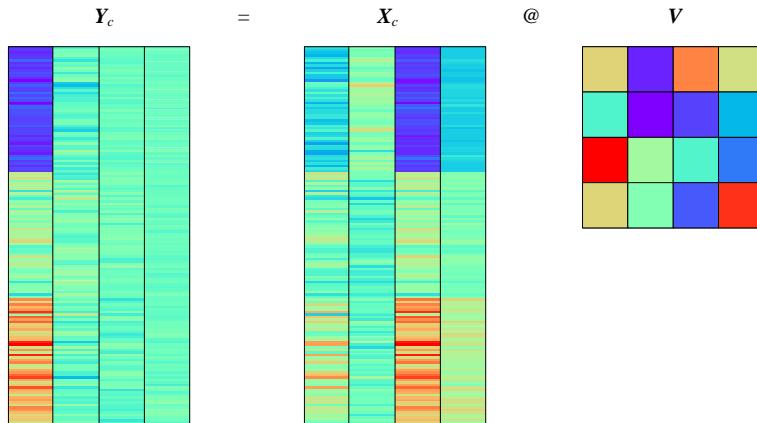
图 17. 矩阵 X_c 进行经济型 SVD 分解

X_c 投影到 V

如图 18 所示，将中心化矩阵 X_c 投影到 V 得到 Y_c ，

$$Y_c = X_c V \quad (39)$$

Y_c 的形状和 X_c 一致。

图 18. 矩阵 X_c 投影到 V

X_c 的质心位于原点， Y_c 的质心也位于原点，即：

$$E(Y_c) = E(X_c V) = E(X_c) V = [0 \ 0 \ 0 \ 0] V = [0 \ 0 \ 0 \ 0] \quad (40)$$

本章前文提过， Y_c 的协方差为：

$$\Sigma_Y = A = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} \quad (41)$$

而原数据矩阵 X 的质心位于 $E(X)$ 。 X_c 和 X 的协方差矩阵完全相同。

几何视角来看， X 到 X_c 是质心从 $E(X)$ 平移到原点。数据本身的分布“形状”相对于质心来说没有任何改变，而协方差矩阵描述的就是分布形状。

X 投影到 V

$V = [v_1, v_2, v_3, v_4]$ 是个 \mathbb{R}^4 规范正交基，不但 X_c 可以投影到 V 中，原始数据 X 也可以投影到 V 中。将 X 投影到 V 得到 Y ，

$$Y = X V \quad (42)$$

Y 的质心显然不在原点， $E(Y)$ 具体位置为：

$$E(Y) = E(X)V = [5.843 \ 3.057 \ 3.758 \ 1.199]V = [5.502 \ -5.326 \ 0.631 \ -0.033] \quad (43)$$

Y 的协方差矩阵则和 Y_c 完全相同，这一点请大家自己证明，并用代码验证。

奇异值 vs 特征值

将 (38) 代入 (6) 得到：

$$\begin{aligned} \Sigma &= \frac{X_c^T X_c}{n-1} = \frac{(USV^T)^T USV^T}{n-1} = \frac{VS^T U^T USV^T}{n-1} \\ &= V \frac{S^2}{n-1} V^T \end{aligned} \quad (44)$$

对比 (44) 和 (20)，可以建立对 Σ 特征值分解和对 X_c 进行 SVD 分解的关系：

$$V \Lambda V^T = V \frac{S^2}{n-1} V^T \quad (45)$$

注意，等式左右两侧的 V 都是正交矩阵，虽然代码计算得到的结果在正负号上会存在差别。

从 (45) 中我们还可以看到 Σ 特征值和 X_c 奇异值之间的量化关系：

$$\underbrace{\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix}}_A = \frac{1}{n-1} \underbrace{\begin{bmatrix} s_1^2 & & & \\ & s_2^2 & & \\ & & \ddots & \\ & & & s_D^2 \end{bmatrix}}_{S^2} \quad (46)$$

即

$$\lambda_j = \frac{1}{n-1} s_j^2 \quad (47)$$

图 19 所示为鸢尾花协方差矩阵特征值和中心化数据奇异值之间的关系。

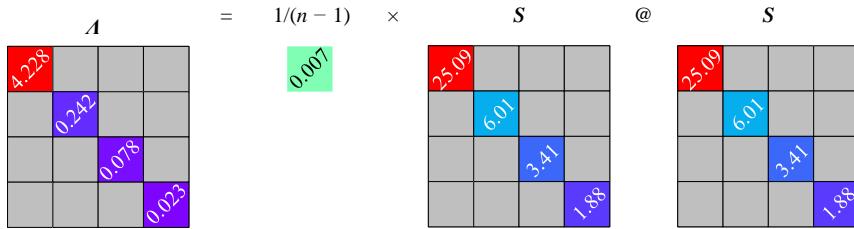


图 19. 特征值和奇异值的关系



有读者可能会问对原数据矩阵 X 直接 SVD 分解，和对 X_c 进行 SVD 分解，两者的区别在哪？这是《数据有道》要探讨的内容。

矩阵乘法第二视角

如图 20 所示，利用矩阵乘法第二视角，(38) 可以展开写成：

$$\begin{aligned} X_c &= \underbrace{\begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_D \end{bmatrix}}_U \underbrace{\begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix}}_S \underbrace{\begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix}}_{V^T} \\ &= s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + s_D \mathbf{u}_D \mathbf{v}_D^T = \sum_{j=1}^D s_j \mathbf{u}_j \mathbf{v}_j^T \end{aligned} \quad (48)$$

同样， \mathbf{u}_j 、 \mathbf{v}_j 仅仅提供投影方向， s_j 决定重要性。

利用向量张量积，上式可以写成：

$$X_c = s_1 \mathbf{u}_1 \otimes \mathbf{v}_1 + s_2 \mathbf{u}_2 \otimes \mathbf{v}_2 + \cdots + s_D \mathbf{u}_D \otimes \mathbf{v}_D = \sum_{j=1}^D s_j \mathbf{u}_j \otimes \mathbf{v}_j \quad (49)$$

这种分解类似图 12。

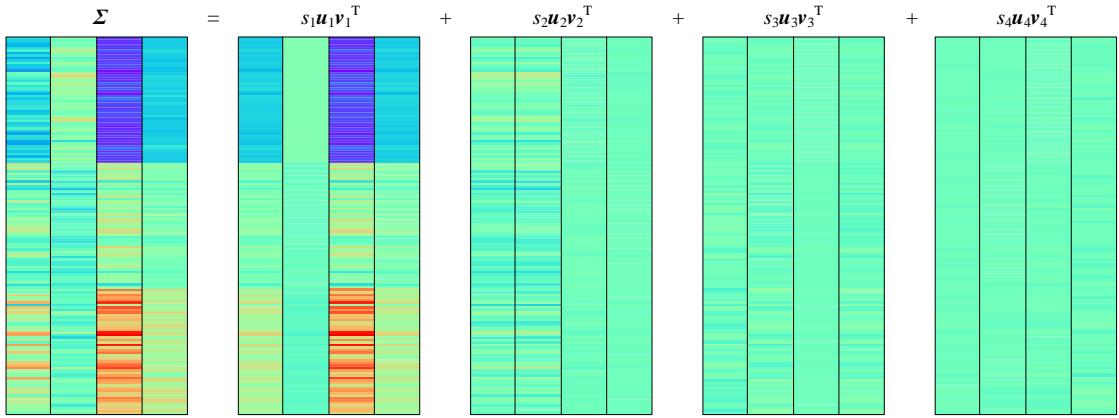


图 20. 利用矩阵乘法第二视角展开 SVD 分解

第二种展开方式

《矩阵力量》第 10 章还介绍过“二次投影”的展开方式，具体如下：

$$\begin{aligned} \mathbf{X}_c = \mathbf{X}_c \mathbf{I} &= \mathbf{X}_c \mathbf{V} \mathbf{V}^T = \mathbf{X}_c \underbrace{\left[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_D \right]}_{\mathbf{V}} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix}_{\mathbf{V}^T} \\ &= \mathbf{X}_c \mathbf{v}_1 \mathbf{v}_1^T + \mathbf{X}_c \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \mathbf{X}_c \mathbf{v}_D \mathbf{v}_D^T = \sum_{j=1}^D \mathbf{X}_c \mathbf{v}_j \mathbf{v}_j^T = \mathbf{X}_c \left(\sum_{j=1}^D \mathbf{v}_j \mathbf{v}_j^T \right) \end{aligned} \quad (50)$$

同样用向量张量积，上式可以写成：

$$\mathbf{X}_c = \mathbf{X}_c \mathbf{v}_1 \otimes \mathbf{v}_1 + \mathbf{X}_c \mathbf{v}_2 \otimes \mathbf{v}_2 + \cdots + \mathbf{X}_c \mathbf{v}_D \otimes \mathbf{v}_D = \mathbf{X}_c \left(\sum_{j=1}^D \mathbf{v}_j \otimes \mathbf{v}_j \right) \quad (51)$$

请大家自行绘制上式的矩阵运算热图。

13.5 Cholesky 分解：列向量坐标

对协方差矩阵 Σ 进行 Cholesky 分解，得到的结果是下三角矩阵 \mathbf{L} 和上三角矩阵 \mathbf{L}^T 乘积：

$$\Sigma = \mathbf{L} \mathbf{L}^T = \mathbf{R}^T \mathbf{R} \quad (52)$$

其中， \mathbf{R} 为上三角矩阵， $\mathbf{R} = \mathbf{L}^T$ 。

图 21 所示为协方差矩阵 Cholesky 分解运算热图。

→ 建议大家回顾《矩阵力量》第 12、24 章，从几何角度、数据角度理解 Cholesky 分解，本节不再重复。

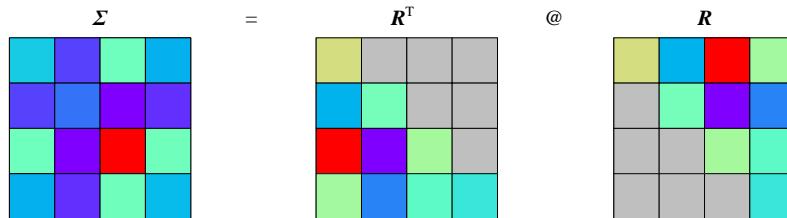


图 21. 对协方差矩阵 Cholesky 分解运算热图

给定数据矩阵 Z , Z 的每个随机变量均服从标准正态分布，且相互独立，也就是 IID; Z 的协方差矩阵为单位矩阵 I ,

$$\Sigma_Z = \frac{Z^T Z}{n-1} = I \quad (53)$$

令

$$X = ZR + E(X) \quad (54)$$

从 (54) 推导 X 的协方差矩阵:

$$\Sigma_X = \frac{(X - E(X))^T (X - E(X))}{n-1} = \frac{(ZR)^T (ZR)}{n-1} = \frac{R^T Z^T ZR}{n-1} = R^T \frac{Z^T Z}{n-1} R = R^T R \quad (55)$$

→ 以上内容对于产生满足特定相关性随机数特别重要，本书第 15 章将展开讲解。

13.6 距离：欧氏距离 vs 马氏距离

协方差矩阵还出现在距离度量运算中，比如马氏距离。本节比较欧氏距离和马氏距离，并引出下一节内容。

欧氏距离

从矩阵运算角度来看，欧氏距离的平方就是《矩阵力量》第 5 章介绍的**二次型** (quadratic form)。比如，空间中任意一点 x 到质心 μ 的欧氏距离为：

$$d^2 = (\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu}) = \|\mathbf{x} - \boldsymbol{\mu}\|_2^2 = \sum_{j=1}^D (x_j - \mu_j)^2 \quad (56)$$

如图 22 (a) 所示，如果 \mathbf{x} 有 2 个特征，即 $D = 2$ ， $d = \|\mathbf{x} - \boldsymbol{\mu}\| = 1$ 代表圆心位于质心 $\boldsymbol{\mu}$ 、半径为 1 的正圆。图 22 (a) 中正圆的解析式为：

$$(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 = 1 \quad (57)$$

如图 22 (b) 所示，如果 \mathbf{x} 有 3 个特征，即 $D = 3$ ， $d = \|\mathbf{x} - \boldsymbol{\mu}\| = 1$ 代表圆心位于质心 $\boldsymbol{\mu}$ 、半径为 1 的正球体，对应的解析式为：

$$(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + (x_3 - \mu_3)^2 = 1 \quad (58)$$

当 $D > 3$ 时， $d = \|\mathbf{x} - \boldsymbol{\mu}\| = 1$ 代表空间中的超球体。

换个角度， $D = 2$ ，当 d 取不同值时，欧氏距离等距线则是一层层同心圆，具体如图 22 (c) 所示。 $D = 3$ ，当 d 取不同值时，欧氏距离等距线变成了一层层同心正球体。

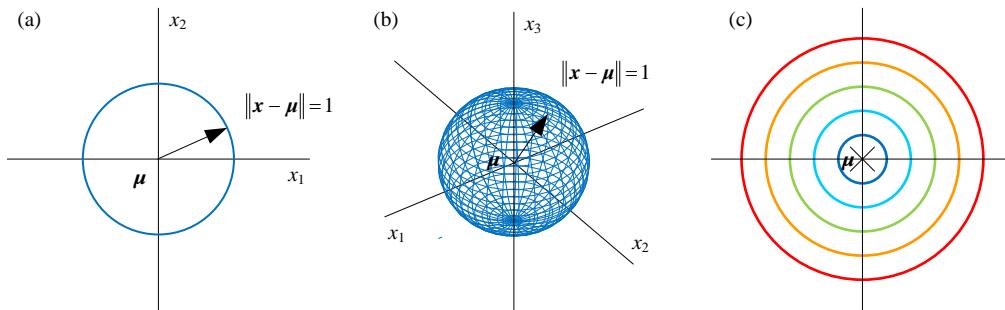


图 22. 正圆、正球体、同心圆

以鸢尾花数据为例，它的质心位于：

$$\boldsymbol{\mu} = \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \quad (59)$$

原点 $\mathbf{0}$ 和质心 $\boldsymbol{\mu}$ 的欧氏距离为：

$$\|\mathbf{0} - \boldsymbol{\mu}\| = \sqrt{(0 - 5.843)^2 + (0 - 3.057)^2 + (0 - 3.758)^2 + (0 - 1.199)^2} \approx 7.684 \quad (60)$$

⚠ 注意，上式中欧氏距离的单位为厘米。

马氏距离

马氏距离的平方也是二次型：

$$d^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \left\| \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu}) \right\|_2^2 \quad (61)$$

如图 23 (a) 所示， $D = 2$ 时， $d = \left\| \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu}) \right\| = 1$ 代表圆心位于质心 $\boldsymbol{\mu}$ 的椭圆。

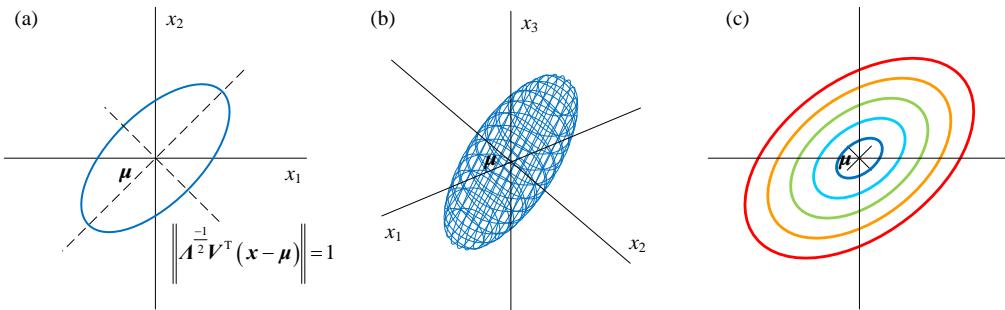


图 23. 椭圆、椭球、同心椭圆

特别地，如果协方差矩阵 $\boldsymbol{\Sigma}$ 为：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \\ & \sigma_2^2 \end{bmatrix}, \quad \sigma_1 > \sigma_2 > 0 \quad (62)$$

马氏距离 $d = 1$ 对应椭圆的解析式为：

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = 1 \quad (63)$$

这个椭圆显然是正椭圆，圆心位于 (μ_1, μ_2) ，半长轴为 σ_1 ，半短轴为 σ_2 。

对于一般的协方差矩阵 $\boldsymbol{\Sigma}_{2 \times 2}$ ，想知道旋转椭圆的半长轴、半短轴长度，则需要利用特征值分解得到其特征值矩阵：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \xrightarrow{\text{EVD}} \mathbf{A} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad (64)$$

这个旋转椭圆的圆心位于 (μ_1, μ_2) ，半长轴为 $\sqrt{\lambda_1}$ ，半短轴为 $\sqrt{\lambda_2}$ 。特征值分解得到的特征向量 \mathbf{v}_1 、 \mathbf{v}_2 则告诉我们椭圆长轴、短轴方向。

如图 22 (b) 所示，如果 \mathbf{x} 有 3 个特征，即 $D = 3$ ， $d = \left\| \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^\top (\mathbf{x} - \boldsymbol{\mu}) \right\| = 1$ 代表圆心位于质心 $\boldsymbol{\mu}$

的椭球体。

同样，如果协方差矩阵 $\boldsymbol{\Sigma}$ 为：

$$\Sigma = \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \sigma_3^2 \end{bmatrix}, \quad \sigma_1 > \sigma_2 > \sigma_3 > 0 \quad (65)$$

马氏距离 $d = 1$ 对应椭球的解析式为：

$$\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} + \frac{(x_3 - \mu_3)^2}{\sigma_3^2} = 1 \quad (66)$$

σ_1 、 σ_2 、 σ_3 都是椭球的半主轴 (principal semi-axis) 长度，我们管它们分别叫第一、第二、第三半主轴长度。

同理，对于更一般的协方差矩阵 $\Sigma_{3 \times 3}$ ，需要通过特征值分解找到半主轴长度 $\sqrt{\lambda_1}$ 、 $\sqrt{\lambda_2}$ 、 $\sqrt{\lambda_3}$ 。三个主轴的方向则分别对应三个特征向量 v_1 、 v_2 、 v_3 。

当 $D > 3$ 时， $d = \left\| A^{\frac{-1}{2}} V^T (x - \mu) \right\| = 1$ 代表空间中的超椭球。

$D = 2$ ，当 d 取不同值时，马氏距离等距线则是一层层同心椭圆，如图 22 (c) 所示。

还是以鸢尾花数据为例，如图 24 所示，原点 θ 和质心 μ 的马氏距离平方值为：

$$d^2 = \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \right)^T \left(\begin{bmatrix} 0.69 & -0.042 & 1.3 & 0.52 \\ -0.042 & 0.19 & -0.33 & -0.12 \\ 1.3 & -0.33 & 3.1 & 1.3 \\ 0.52 & -0.12 & 1.3 & 0.58 \end{bmatrix} \right)^{-1} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \right) = 129.245 \quad (67)$$

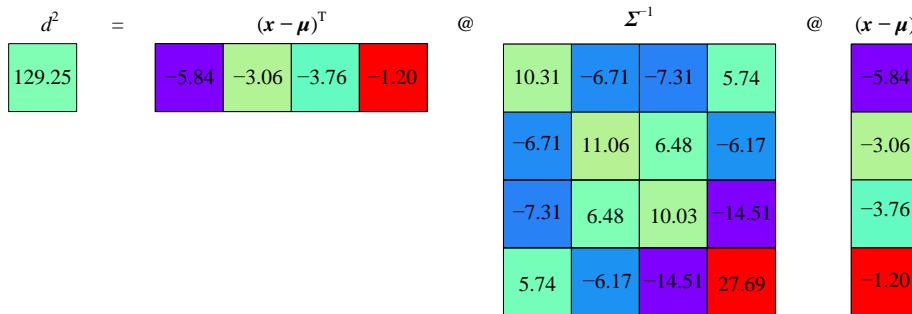


图 24 计算 d^2 的矩阵运算热图

(67) 开平方得到原点 θ 和质心 μ 的马氏距离为：

$$d = \sqrt{129.245} = 11.3686 \quad (68)$$

马氏距离没有单位。更准确地说，马氏距离的单位是标准差，比如 $d = 11.3686$ 代表马氏距离为“11.3686 个均方差”。



本书第 23 章还会继续探讨马氏距离。

有了本节内容铺垫，下一节深入探讨协方差的几何内涵。



Bk5_Ch13_01.py 绘制本章前文大部分矩阵运算热图。

13.7 几何视角：超椭球、椭球、椭圆

“旋转” 超椭球

根据上一节所学，如果 $D = 4$, $(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = 1$ 代表四维空间 \mathbb{R}^4 圆心位于 $\boldsymbol{\mu}$ 的超椭球。我们知道，对于鸢尾花样本数据 \mathbf{X} ，在 \mathbb{R}^4 中代表数据的超椭球的圆心位于 $E(\mathbf{X})$ ，即：

$$E(\mathbf{X}) = [5.843 \quad 3.057 \quad 3.758 \quad 1.199] \quad (69)$$

根据图 10 中所示的对 $\boldsymbol{\Sigma}$ 特征值分解，我们知道超椭球的四个半主轴长度分别为：

$$\begin{aligned}\sqrt{\lambda_1} &\approx \sqrt{4.228} \approx 2.056 \text{ cm} \\ \sqrt{\lambda_2} &\approx \sqrt{0.242} \approx 0.492 \text{ cm} \\ \sqrt{\lambda_3} &\approx \sqrt{0.078} \approx 0.279 \text{ cm} \\ \sqrt{\lambda_4} &\approx \sqrt{0.023} \approx 0.154 \text{ cm}\end{aligned} \quad (70)$$

\mathbb{R}^4 中超椭球四个主轴所在方向对应图 10 中 \mathbf{V} 的四个列向量，即：

$$\mathbf{V} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \mathbf{v}_3 \quad \mathbf{v}_4] = \begin{bmatrix} 0.751 & 0.284 & 0.502 & 0.321 \\ 0.380 & 0.547 & -0.675 & -0.317 \\ 0.513 & -0.709 & -0.059 & -0.481 \\ 0.168 & -0.344 & -0.537 & 0.752 \end{bmatrix} \quad (71)$$

显然，在纸面上很难可视化一个四维空间的超椭球，因此我们选择用投影的办法将超椭球投影在不同三维空间、二维平面上。

“旋转” 超椭球投影到三维空间

图 25 (a) 所示为四维空间超椭球在 $x_1x_2x_3$ 这个三维空间的投影，结果是个圆心位于质心的椭球。

为了获得这个椭球的解析式，我们先将 4×4 协方差矩阵 Σ “投影”到图 25 (a) 这个三维空间中，我们把这个新的协方差记做：

$$\Sigma_{1,2,3} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} \underbrace{\begin{bmatrix} 0.686 & -0.042 & 1.274 & 0.516 \\ -0.042 & 0.190 & -0.330 & -0.122 \\ 1.274 & -0.330 & 3.116 & 1.296 \\ 0.516 & -0.122 & 1.296 & 0.581 \end{bmatrix}}_{\Sigma} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = \begin{bmatrix} 0.686 & -0.042 & 1.274 \\ -0.042 & 0.190 & -0.330 \\ 1.274 & -0.330 & 3.116 \end{bmatrix} \quad (72)$$

Σ 消去了第 4 行和第 4 列得到 $\Sigma_{1,2,3}$ 。

从数据角度来看，原始数据矩阵 $X_{150 \times 4}$ 先投影得到 $X_{1,2,3}$ ：

$$X_{1,2,3} = \underbrace{\begin{bmatrix} x_1 & x_2 & x_3 & x_4 \end{bmatrix}}_X \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \quad (73)$$

如上运算相当于，保留了 X 的前三列数据。 $X_{1,2,3}$ 再算协方差矩阵结果就是 $\Sigma_{1,2,3}$ 。

单位矩阵 $I_{4 \times 4}$ 是 \mathbb{R}^4 的标准正交系，可以写成：

$$I_{4 \times 4} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = [\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3 \ \mathbf{e}_4] \quad (74)$$

(72) 相当于 X 在 $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$ 基底中投影。

四维空间的超椭球的圆心 $E(X)$ 在图 25 (a) 这个三维空间的位置很容易计算：

$$E(X)[\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3] = [5.843 \ 3.057 \ 3.758 \ 1.199] \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{bmatrix} = [5.843 \ 3.057 \ 3.758] \quad (75)$$

如果想要调换图 25 (a) 中 x_1 和 x_2 的顺序，只需要 $[\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3]$ 乘上如下的置换矩阵 (permutation matrix)：

$$[\mathbf{e}_1 \ \mathbf{e}_2 \ \mathbf{e}_3] \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix} = [\mathbf{e}_2 \ \mathbf{e}_1 \ \mathbf{e}_3] \quad (76)$$



《矩阵力量》第 5 章讲过置换矩阵，大家可以回顾。

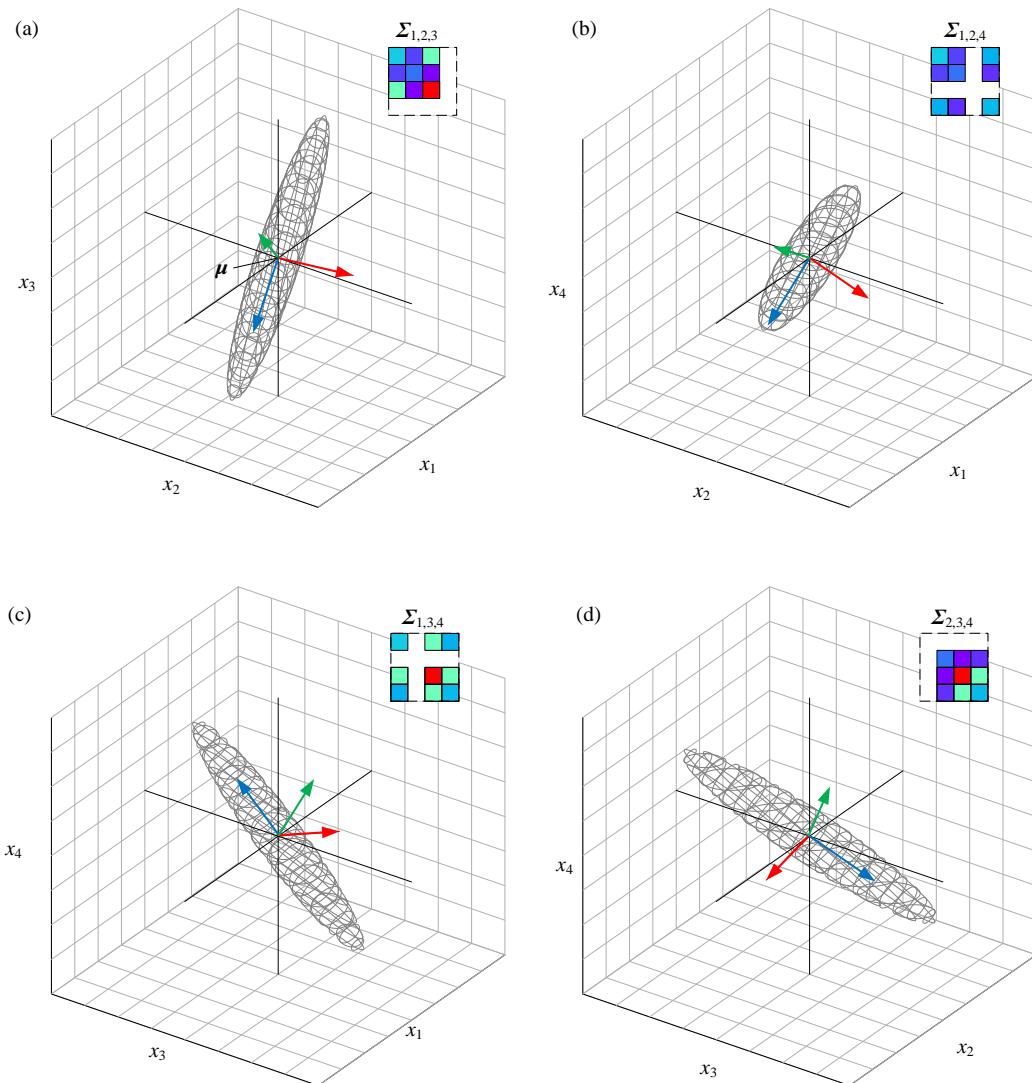


图 25 四维空间的“旋转”超椭球在三维空间中的四个投影

图 25 (a) 中的蓝、红、绿箭头分别代表三维椭球的第一、第二、第三主轴方向。这三个主轴方向需要特征值分解 (72) 中协方差矩阵：

$$\begin{aligned} \Sigma_{1,2,3} &= \begin{bmatrix} 0.686 & -0.042 & 1.274 \\ -0.042 & 0.190 & -0.330 \\ 1.274 & -0.330 & 3.116 \end{bmatrix} \\ &= \begin{bmatrix} -0.389 & 0.662 & 0.639 \\ 0.091 & -0.663 & 0.743 \\ -0.916 & -0.347 & -0.198 \end{bmatrix} \begin{bmatrix} 3.691 & & \\ & 0.059 & \\ & & 0.241 \end{bmatrix}^T \end{aligned} \quad (77)$$

由此，我们知道图 25 (a) 中椭球的三个半主轴的长度为：

$$\begin{aligned}\sqrt{3.691} &\approx 1.921 \text{ cm} \\ \sqrt{0.059} &\approx 0.243 \text{ cm} \\ \sqrt{0.241} &\approx 0.491 \text{ cm}\end{aligned}\tag{78}$$

(77) 的特征值分解也帮我们求得椭球的三个主轴方向。

注意，图 25 (a) 中的蓝、红、绿箭头显然不是 (71) 中 \mathbf{V} 在 \mathbb{R}^3 中投影，原因很简单 \mathbf{V} 在 \mathbb{R}^3 中应该有四个“影子”，而不是三个。这一点在图 26 中看得更明显。

只有 \mathbf{V} 在沿着 \mathbf{v}_j 方向投影（注意，不是在 \mathbf{v}_j 方向投影）， \mathbf{v}_j 的分量才会消失。这就好比，正午阳光下，一根柱子相当于“没有”影子。

请大家自行分析图 25 剩余三幅子图，并写出对应的投影运算。

“旋转” 椭球投影到二维平面

图 26 所示为图 25 (a) 中椭球进一步投影到三个二维平面上。

以 x_1x_2 平面为例，先将 4×4 协方差矩阵 Σ 投影 x_1x_2 平面，结果为：

$$\Sigma_{1,2} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & \\ & & & \end{bmatrix} \underbrace{\begin{bmatrix} 0.686 & -0.042 & 1.274 & 0.516 \\ -0.042 & 0.190 & -0.330 & -0.122 \\ 1.274 & -0.330 & 3.116 & 1.296 \\ 0.516 & -0.122 & 1.296 & 0.581 \end{bmatrix}}_{\Sigma} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & \\ & & & \end{bmatrix} = \begin{bmatrix} 0.686 & -0.042 \\ -0.042 & 0.190 \end{bmatrix} \tag{79}$$

请大家自己写出数据投影对应的矩阵运算。

为了计算 (79) 协方差对应的椭圆，需要对其特征值分解：

$$\Sigma_{1,2} = \begin{bmatrix} 0.686 & -0.042 \\ -0.042 & 0.190 \end{bmatrix} = \begin{bmatrix} 0.996 & 0.084 \\ -0.084 & 0.996 \end{bmatrix} \begin{bmatrix} 0.689 & & \\ & 0.186 & \\ & & \end{bmatrix} \begin{bmatrix} 0.996 & 0.084 \\ -0.084 & 0.996 \end{bmatrix}^T \tag{80}$$

通过上述特征值分解，我们知道在 x_1x_2 平面上椭圆的半长轴、半短轴长度分别为 0.830、0.431。单位都是厘米 cm。

此外，请大家注意图 25 (a) 中 x_1x_2 平面上这个椭圆中背景蓝色的矩形，这是本节后续要讨论的内容。

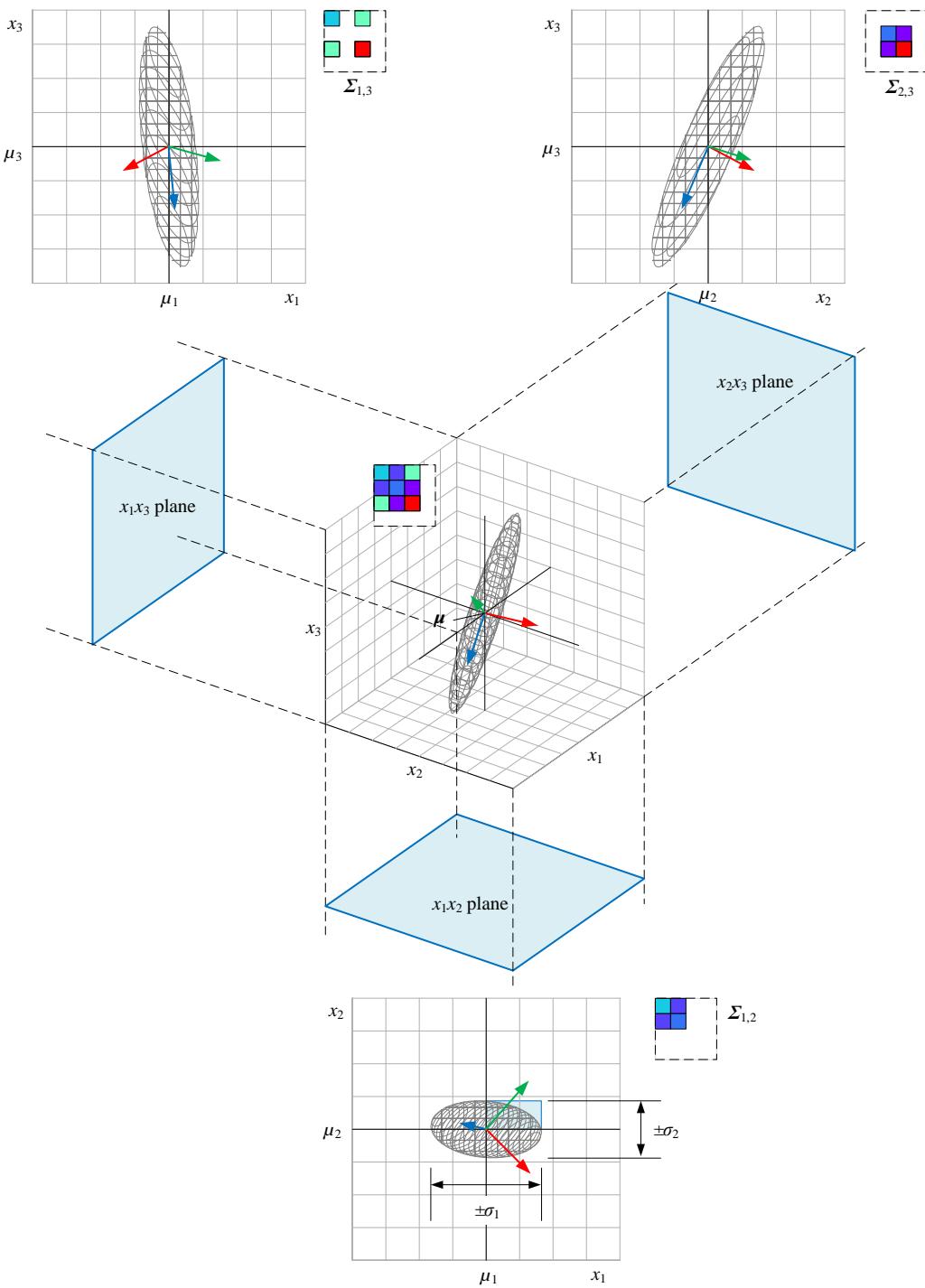


图 26. “旋转”椭球投影到三个二维平面

不考虑 x_i 、 x_j ($i \neq j$) 顺序的话， \mathbb{R}^4 中超椭球朝 x_ix_j 面投影，一共可以获得 6 个不同平面上的椭圆投影结果，具体如图 27 所示。请大家自行分析图 27 中这 6 幅子图。

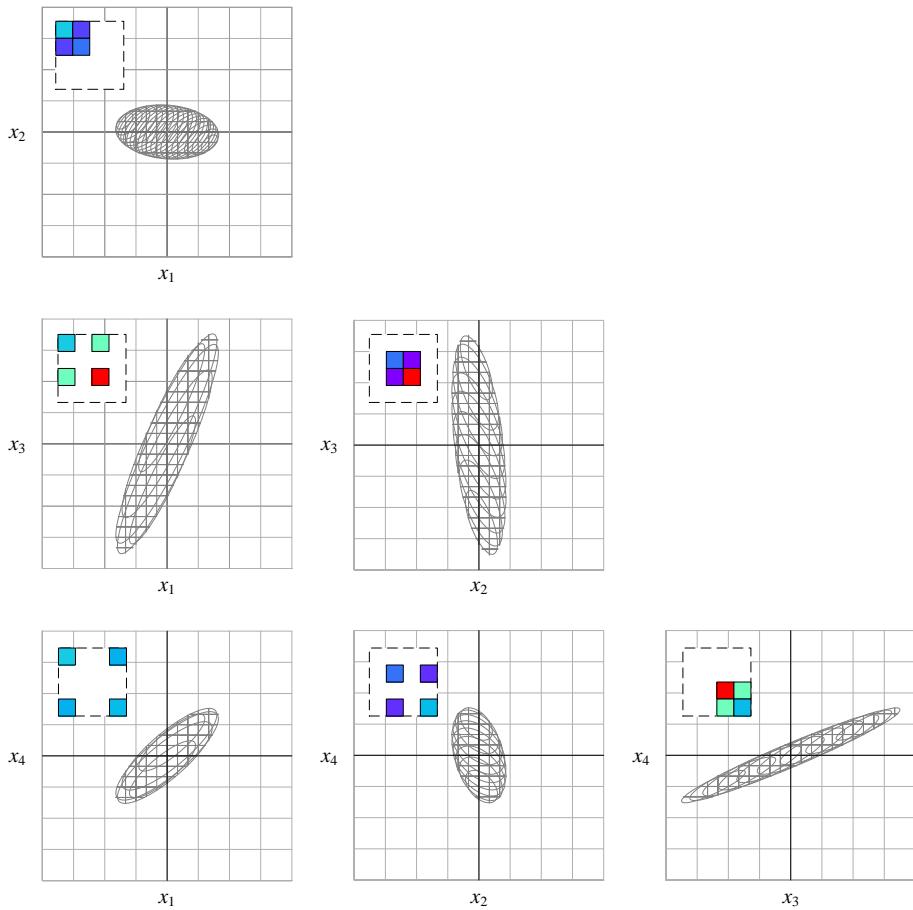


图 27. “旋转”超椭球在 6 个平面上的投影结果

矩形的面积、对角线长度

如图 28 (a) 所示，椭圆相切于矩形的四条边。该矩形的四个顶点分别是 $(\mu_1 - \sigma_1, \mu_2 - \sigma_2)$ 、 $(\mu_1 - \sigma_1, \mu_2 + \sigma_2)$ 、 $(\mu_1 + \sigma_1, \mu_2 + \sigma_2)$ 、 $(\mu_1 + \sigma_1, \mu_2 - \sigma_2)$ 。

图 28 (c) 中矩形的四个顶点分别为 (μ_1, μ_2) 、 $(\mu_1, \mu_2 + \sigma_2)$ 、 $(\mu_1 + \sigma_1, \mu_2 + \sigma_2)$ 、 $(\mu_1 + \sigma_1, \mu_2)$ 。

图 28 (a) 所示矩形的面积为 $4\sigma_1\sigma_2$ ，而图 28 (c) 中矩形为图 28 (a) 矩形的 $1/4$ ，对应面积为 $\sigma_1\sigma_2$ 。

图 28 (c) 中 $1/4$ 矩形对角线长度为 $\sqrt{\sigma_1^2 + \sigma_2^2}$ ，这个值是其协方差迹的平方根，即：

$$\sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\text{tr}(\Sigma_{2 \times 2})} \quad (81)$$

图 28 (b) 所示矩形也和椭圆相切于四条边，两组对边分别平行于 v_1 、 v_2 。这个矩形的面积为 $4\sqrt{\lambda_1\lambda_2}$ 。而图 28 (d) 中矩形为图 28 (b) 矩形的 $1/4$ ，对应面积为 $\sqrt{\lambda_1\lambda_2}$ 。

$\sqrt{\lambda_1\lambda_2}$ 是协方差行列式值的平方根：

$$\sqrt{\lambda_1 \lambda_2} = \sqrt{|\Lambda_{2 \times 2}|} = \sqrt{|\Sigma_{2 \times 2}|} \quad (82)$$

图 28 (d) 中 1/4 矩形对角线长度为 $\sqrt{\lambda_1 + \lambda_2}$ ，和图 28 (c) 中矩形对角线长度相同，即：

$$\sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\text{tr}(\Sigma_{2 \times 2})} = \sqrt{\text{tr}(\Lambda_{2 \times 2})} = \sqrt{\lambda_1 + \lambda_2} \quad (83)$$

这是本书下一章要讨论协方差的重要几何性质之一。

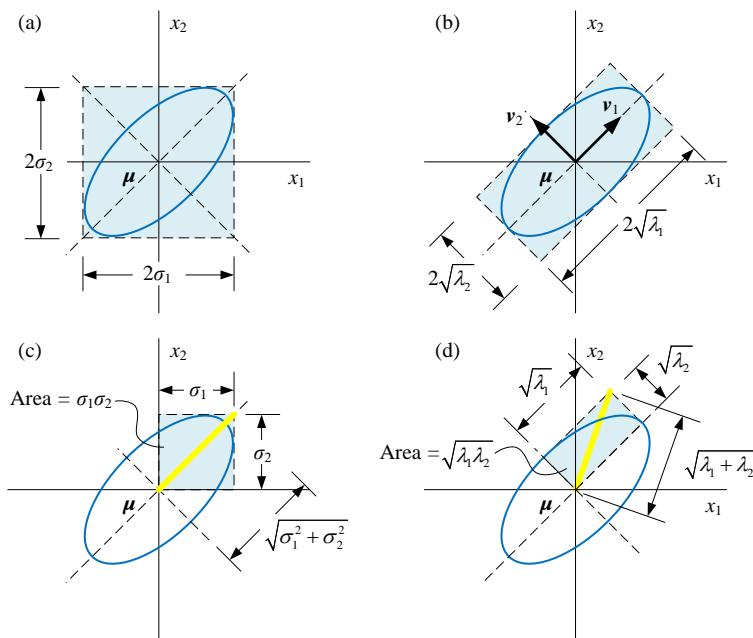


图 28. 和椭圆相切矩形的面积

“正”超椭球投影到三维空间

本节前文的“旋转”超椭球经过旋转之后得到“正”超椭球，这个“正”超椭球对应的协方差矩阵为 Λ ，具体值为：

$$\Lambda = \begin{bmatrix} 4.228 & & & \\ & 0.242 & & \\ & & 0.078 & \\ & & & 0.023 \end{bmatrix} \quad (84)$$

这个“正”超椭球的解析式为：

$$\frac{y_1^2}{4.228} + \frac{y_2^2}{0.242} + \frac{y_3^2}{0.078} + \frac{y_4^2}{0.023} = 1 \quad (85)$$

图 29 所示为“正”超椭球在四个三维空间中投影得到的椭球。其中，图 29 (a) 所示为“正”超椭球在 $y_1 y_2 y_3$ 这个三维空间的投影，对应的解析式为：

$$\frac{y_1^2}{4.228} + \frac{y_2^2}{0.242} + \frac{y_3^2}{0.078} = 1 \quad (86)$$

图 29 (a) 中蓝、红、绿色箭头对应为上述“正”超椭球的第一、第二、第三主轴方向。

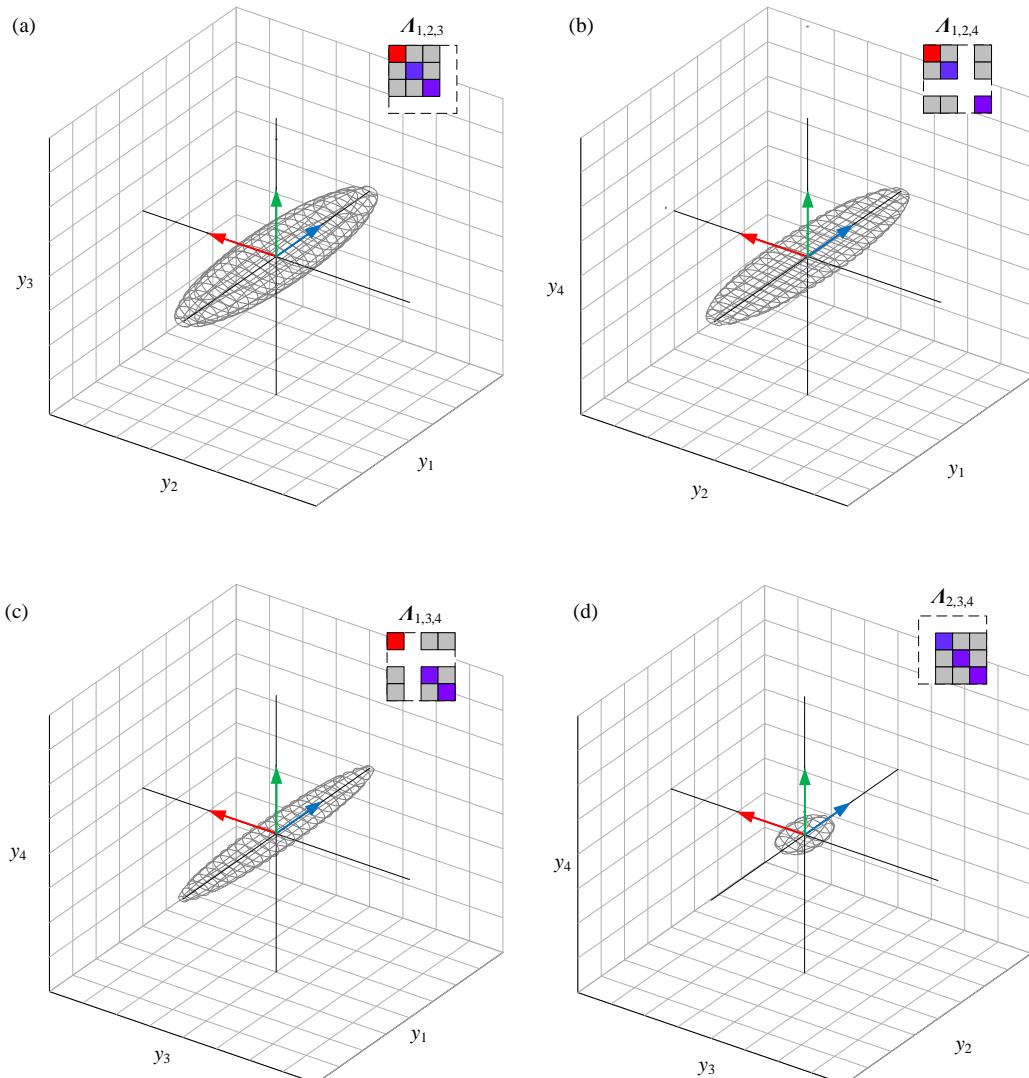


图 29 四维空间的“正”超椭球在三维空间中的四个投影

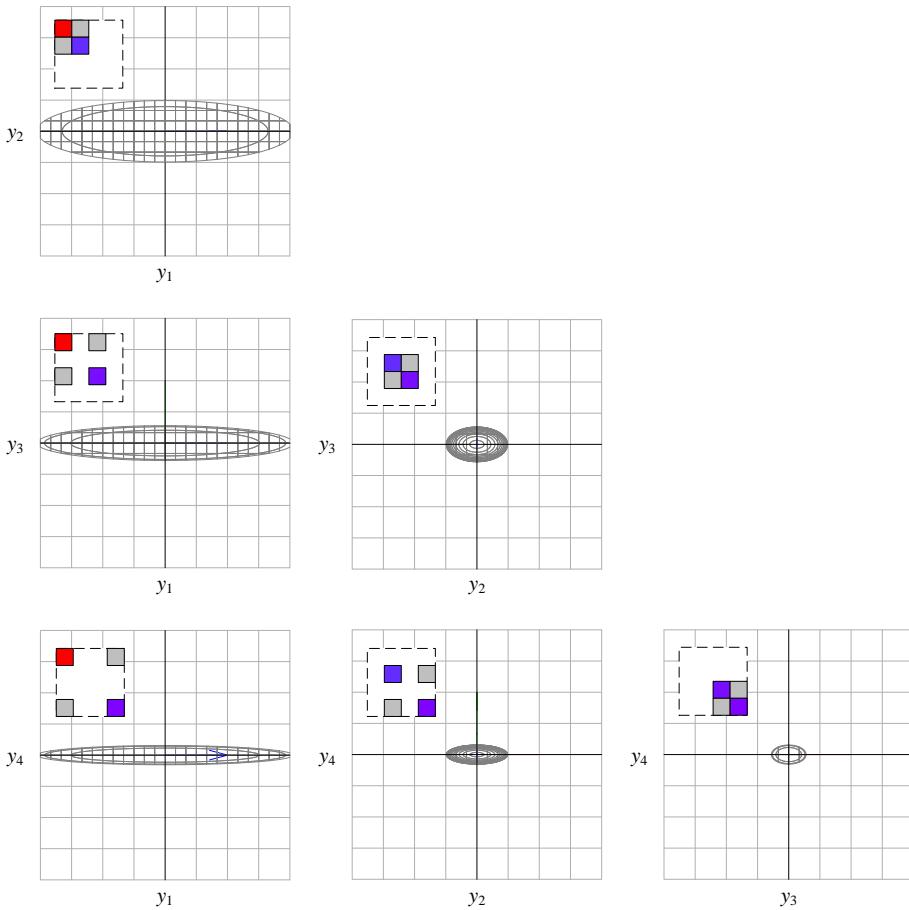


图 30. “正”超椭球在 6 个平面上的投影结果

相关系数矩阵

大家是否立刻想到，相关系数矩阵 P 也可以做特征值分解，也就是说 P 也可以有类似前文协方差矩阵的几何解释。

根据图 16，相关系数矩阵 P 对应的超椭球的半主轴长度分别为 $\sqrt{2.918} = 1.708$ 、
 $\sqrt{0.914} = 0.956$ 、 $\sqrt{0.146} = 0.383$ 、 $\sqrt{0.021} = 0.143$ 。

图 31 所示为相关系数矩阵所代表的四维空间的“旋转”超椭球在三维空间中的四个投影。图 32 所示为这个超椭圆在六个平面的投影。请大家自行分析这两幅图，特别是方差、标准差。

注意，相关系数矩阵可以视作 Z 分数的协方差矩阵。

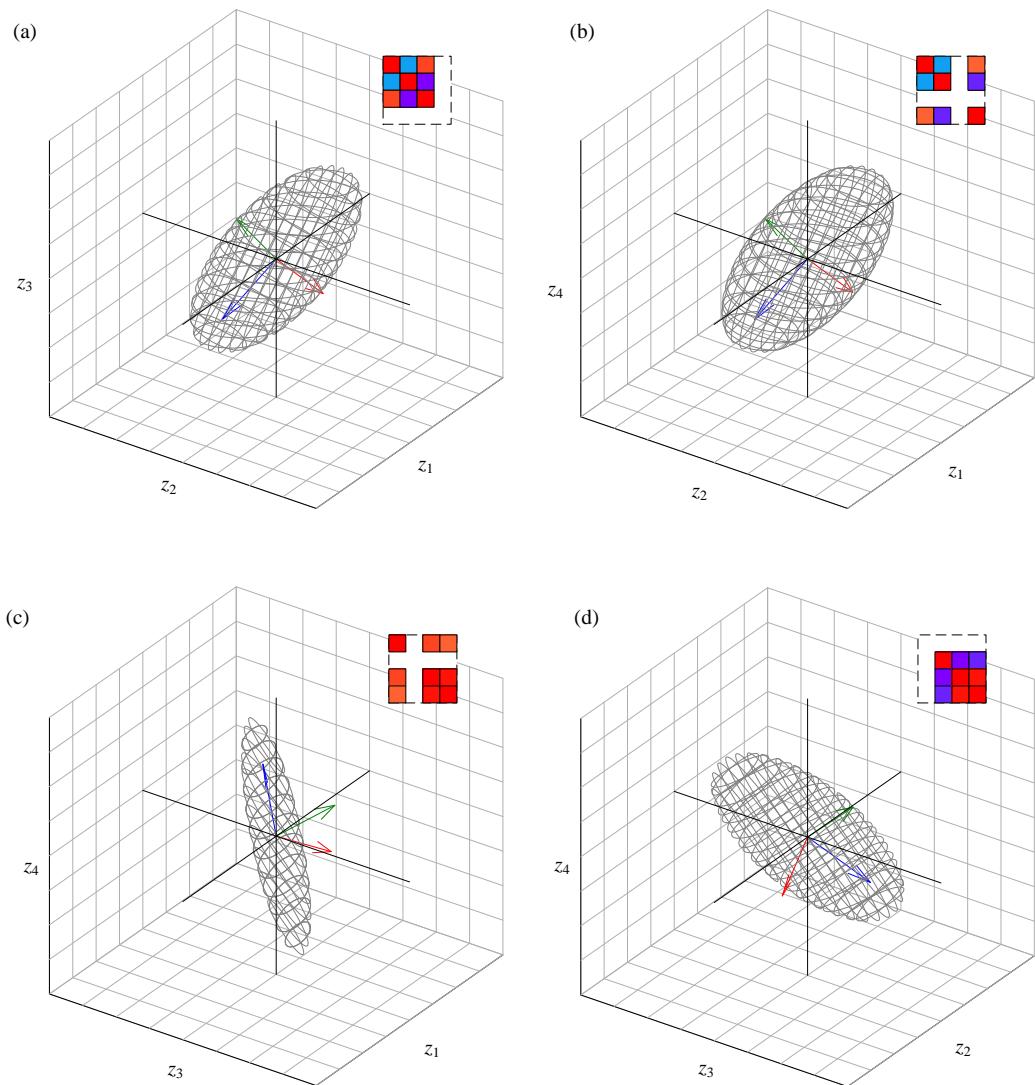


图 31 四维空间的“旋转”超椭球在三维空间中的四个投影，相关系数矩阵

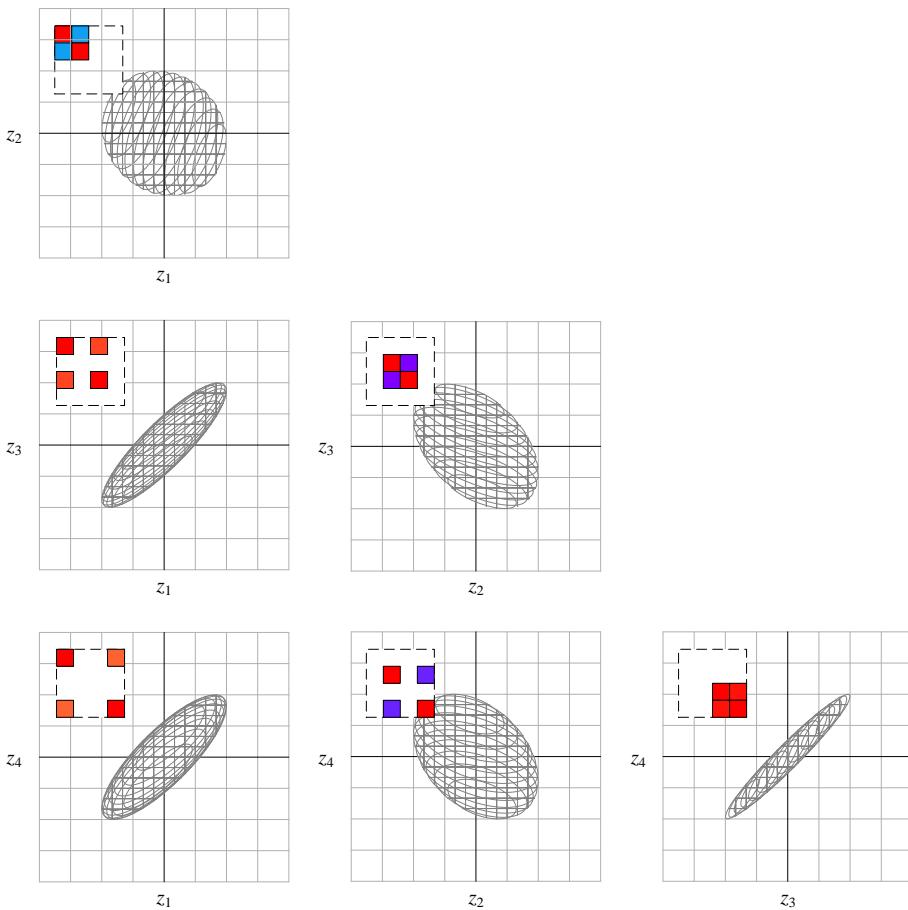


图 32. "旋转"超椭球在 6 个平面上的投影结果，相关性系数矩阵

13.8 合并协方差矩阵

本节介绍一个概念——**合并协方差矩阵** (pooled covariance matrix)，定义为：

$$\boldsymbol{\Sigma}_{\text{pooled}} = \frac{1}{\sum_{k=1}^K (n_k - 1)} \sum_{k=1}^K (n_k - 1) \boldsymbol{\Sigma}_k = \frac{1}{n - K} \sum_{k=1}^K (n_k - 1) \boldsymbol{\Sigma}_k \quad (87)$$

其中，\$n\$ 代表总体样本数，\$n_k\$ 为标签为 \$C_k\$ 的样本数，\$K\$ 为标签数量。\$\boldsymbol{\Sigma}_i\$ 是标签为 \$C_k\$ 的样本数据协方差矩阵。

上式相当于加权平均，这么做是为了保证整体协方差矩阵的无偏性，因为每个组内的样本数可能不同，直接将所有样本合并起来计算协方差矩阵可能会导致估计偏差。

如果假设分类质心重叠，合并协方差矩阵可以用来估算样本整体方差。合并协方差矩阵可以用来比较不同子集的协方差之间的差异，也就是不同分类标签数据的分布情况。此外，我们会在“鸢尾花书”《数据有道》主成分分析中看到合并协方差的应用。

以鸢尾花数据矩阵为例，总体样本数为 $n = 150$ ，一共有三种 ($K = 3$) 标签 C_1 、 C_2 、 C_3 ，分别对应的样本数为 $n_1 = 50$ 、 $n_2 = 50$ 、 $n_3 = 50$ 。合并协方差矩阵为：

$$\Sigma_{\text{pooled}} = \frac{1}{150-3} \sum_{k=1}^3 (50-1) \Sigma_k = \frac{49}{147} \times (\Sigma_1 + \Sigma_2 + \Sigma_3) \quad (88)$$

图 33 中三个彩色的椭圆代表 Σ_1 、 Σ_2 、 Σ_3 ，对应马氏距离为 1。

注意，图 33 中并没有展示 Σ_{pooled} 。

图 33 中 Σ 代表整体数据协方差矩阵。 Σ 完全不同于 Σ_{pooled} 。也可以说， Σ_{pooled} 只是 Σ 的一部分。 Σ_{pooled} 仅仅考虑标签子集数据的协方差矩阵，没有考虑子集之间的分布差异（分类质心的差异）。因此，图 33 中 Σ 对应的旋转椭圆远大于 Σ_1 、 Σ_2 、 Σ_3 。

换个角度来看，合并协方差矩阵相当于，全方差定理中的条件方差的期望，缺少的成分是条件期望的方差。

为了方便比较不同分类协方差矩阵，我们可以将所有椭圆中心重合，得到图 34。 Σ_{pooled} 的对应图 34 中的黑色划线椭圆。比较彩色椭圆和黑色划线椭圆，可以知道不同标签数据分布之间差异。

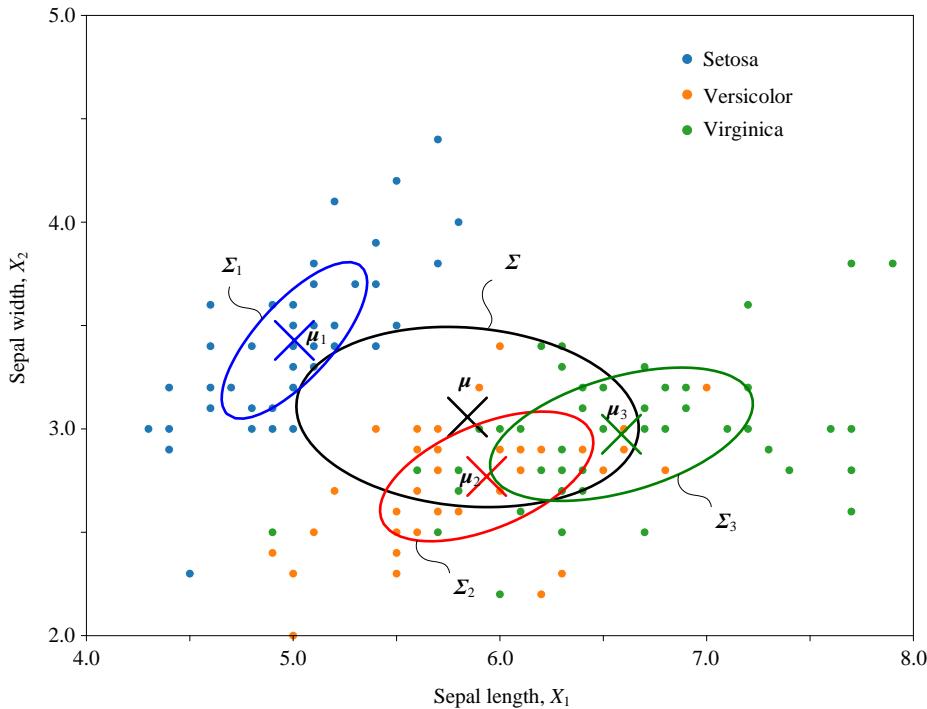
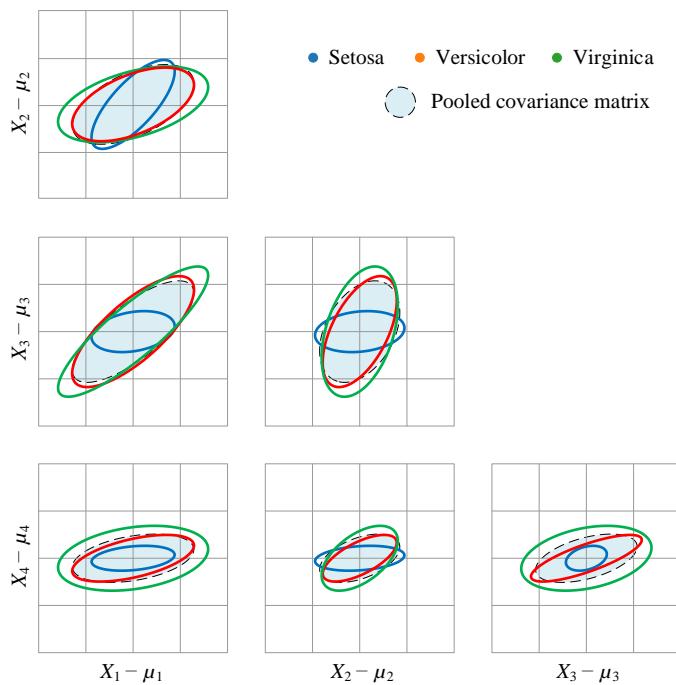


图 33. 分类协方差矩阵、整体协方差矩阵马氏距离为 1 椭圆，花萼长度、花萼宽度

图 34. 马氏距离 1 椭圆, Σ_1 、 Σ_2 、 Σ_3 和合并协方差矩阵 Σ_{pooled} 

这一章结束了本书“高斯”这一板块。这个板块以高斯分布为主线，分别介绍了一元、二元、多元、条件高斯分布，最后介绍了多元高斯分布中的主角——协方差矩阵。相信通过这几章的学习，大家已经看到了线性代数工具在多元统计中的重要作用。

多元统计数据通常表示为向量或矩阵形式，线性代数提供了处理和计算这些对象的基本工具。例如，我们可以使用矩阵运算来计算协方差矩阵、进行线性变换、求解线性方程组等。

在多元统计中，特征值和特征向量是非常重要的概念。通过计算特征值和特征向量，我们可以识别出数据中的主要方向和结构，从而进行降维、聚类、分类等任务。

奇异值分解被广泛用于主成分分析 (PCA)、矩阵分解、压缩和图像处理等任务中。此外，我们可以使用特征值分解或奇异值分解来分析数据的主要结构和变化模式，使用矩阵迹、行列式等概念来计算协方差矩阵的性质，使用矩阵乘法、转置等运算来进行矩阵变换等。

在多元统计中，很多问题可以被视为一个优化问题。线性代数提供了很多优化方法和技巧，例如梯度下降、牛顿法、共轭梯度法等，可以用来解决最小化误差、最大化似然等问题。大家会在本书后续看到更多线性代数在多元统计、数据分析、机器学习领域的应用。



协方差估计的方法还有很多，请大家参考：

<https://scikit-learn.org/stable/modules/covariance.html>

有关合并协方差矩阵，请大家参考：

<https://arxiv.org/pdf/1805.05756.pdf>

14

Functions of Random Variables

随机变量的函数

从几何视角探讨随机变量的线性变换



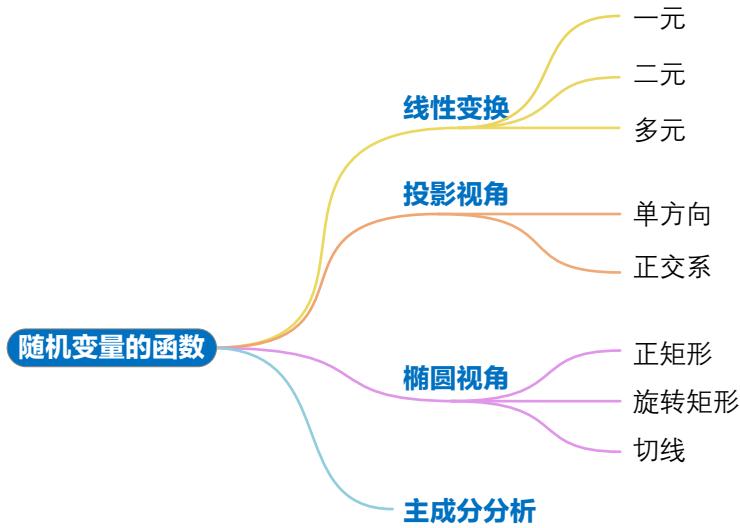
自然的一般规律在大多数情况下不是直接的感知对象。

The general laws of Nature are not, for the most part, immediate objects of perception.

—— 乔治·布尔 (George Boole) | 英格兰数学家和哲学家 | 1815 ~ 1864



- ▶ `numpy.cov()` 计算协方差矩阵
- ▶ `numpy.linalg.eig()` 特征值分解
- ▶ `numpy.linalg.svd()` 奇异值分解
- ▶ `sklearn.decomposition.PCA()` 主成分分析函数
- ▶ `seaborn.heatmap()` 绘制热图
- ▶ `seaborn.kdeplot()` 绘制 KDE 核概率密度估计曲线
- ▶ `seaborn.pairplot()` 绘制成对分析图



14.1 随机变量的函数：以鸢尾花为例

随机变量的函数可以分为两类：**线性变换** (linear transformation)、**非线性变换** (nonlinear transformation)。线性变换是本章的核心内容。

我们在本书第3、4章聊过色子点数的“花式玩法”，比如点数之和、点数平均值、点数之差、点数平方、点数之商等等。这些“花式玩法”都可以叫做随机变量的函数。

比如，点数之和 ($X_1 + X_2$)、点数之差 ($X_1 - X_2$)、点数平均值 ($(X_1 + X_2)/2$) 等都是线性变换。此外，去均值 ($X_1 - E(X_1)$)、标准化 ($(X_1 - E(X_1))/\text{std}(X_1)$) 也都是常见的随机变量的线性变换。

线性变换之外的随机变量变换都统称为非线性变换，比如平方 (X_1^2)、平方求和 ($\sum_j X_j^2$)、乘积 ($X_1 X_2$)、比例 (X_1/X_2)、倒数 ($1/X_1$)、对数变换 ($\ln X_1$) 等等。此外，本书第9章介绍的经验分布累积函数 ECDF 也是常用的非线性变换，ECDF 将原始数据转化成 (0, 1) 区间之内的分位值。

⚠ 注意，经过转换后的随机变量，其分布类型、期望、方差等都可能会发生变化。

从数据角度来看，以上变换又叫**数据转化** (data transformation)，这是《数据有道》一册的话题。

以鸢尾花数据为例

鸢尾花数据的前4列特征分别为花萼长度 (X_1)、花萼宽度 (X_2)、花瓣长度 (X_3)、花瓣宽度 (X_4)。假如在一个有关鸢尾花的研究中，为了进一步挖掘鸢尾花数据中可能存在的量化关系，我们可以分析如下几个指标：

- ▶ 花萼长度去均值，即 $X_1 - E(X_1)$ ；
- ▶ 花萼宽度去均值，即 $X_2 - E(X_2)$ ；
- ▶ 花萼长度、宽度之和，即 $X_1 + X_2$ ；
- ▶ 花萼长度、宽度之差，即 $X_1 - X_2$ ；
- ▶ 花萼长度、宽度乘积，即 $X_1 X_2$ ；
- ▶ 花萼长度、宽度比例，即 X_1/X_2 。

图1所示为经过上述转换后得到的鸢尾花新特征之间的成对特征散点图。这些新特征之间的成对关系中，有些展现出明显的线性关系，有些特征更方便判别鸢尾花分类，有些特征展现出更好的“正态性”，有些则更容易发现“离群值”。

请大家利用成对特征图分析更多鸢尾花特征的随机变量函数。此外，请大家依照同样的方法分析花瓣长度、宽度数据，并且交叉分析花萼、花瓣量化关系。

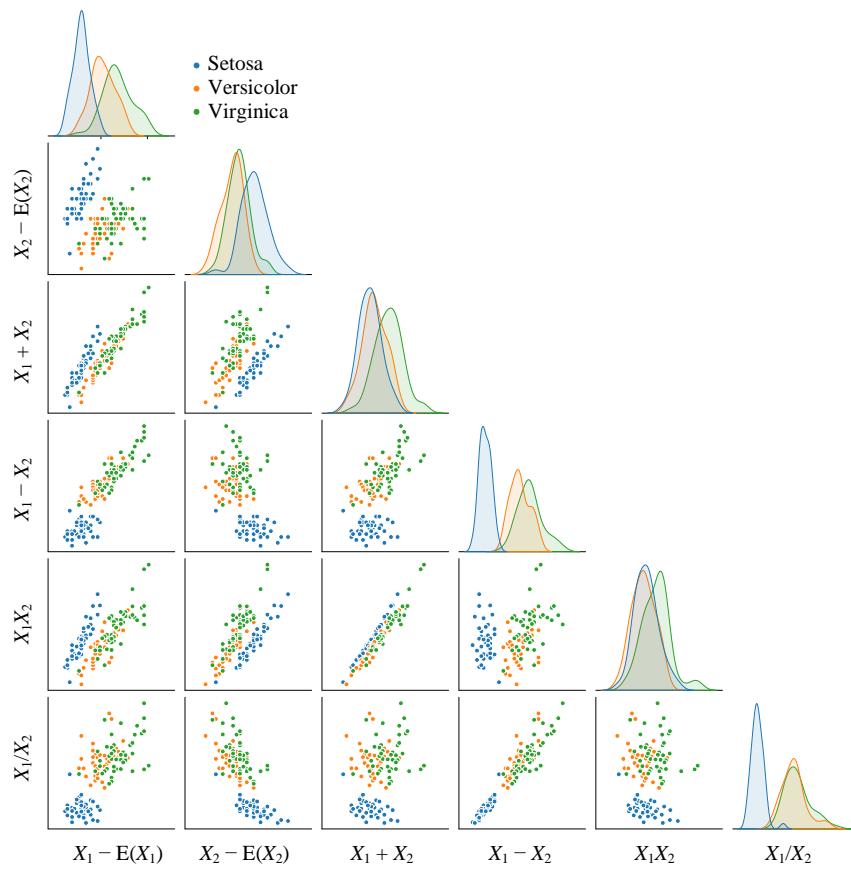


图 1. 鸢尾花花萼长度、宽度特征完成转换后的成对特征散点图

14.2 线性变换：投影视角

《矩阵力量》第 25 章介绍过随机变量的线性变换，我们部分内容“抄”过来。本章后文会用鸢尾花数据展开讲解。

一元随机变量

如果 X 为一个随机变量，对 X 进行函数变换，可以得到其他的随机变量 Y ：

$$Y = h(X) \quad (1)$$

特别地，如果 $h()$ 为线性函数，则 X 到 Y 进行的就是线性变换，比如：

$$Y = h(X) = aX + b \quad (2)$$

其中， a 和 b 为常数。

上式相当于几何中的缩放、平移两步操作。在线性代数中，上式相当于**仿射变换** (affine transformation)。

展开来说，在线性代数中，仿射变换是指一类在二维或三维欧几里得空间中的变换，可以描述为一种线性变换和一个平移向量的组合。与仿射变换不同，线性变换仅由矩阵乘法表达，它可用来缩放、旋转、镜像、剪切一个图形，但不能进行平移操作。

(2) 中， Y 的期望和 X 的期望之间关系：

$$\mathbb{E}(Y) = a\mathbb{E}(X) + b \quad (3)$$

(2) 中， Y 和 X 方差之间关系：

$$\text{var}(Y) = \text{var}(aX + b) = a^2 \text{var}(X) \quad (4)$$

二元随机变量

如果 Y 和二元随机变量 (X_1, X_2) 存在如下关系：

$$Y = aX_1 + bX_2 \quad (5)$$

(5) 可以写成：

$$Y = [a \ b] \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (6)$$

Y 和二元随机变量 (X_1, X_2) 期望之间存在如下关系：

$$\mathbb{E}(Y) = \mathbb{E}(aX_1 + bX_2) = a\mathbb{E}(X_1) + b\mathbb{E}(X_2) \quad (7)$$

(7) 可以写成如下矩阵运算形式：

$$\mathbb{E}(Y) = [a \ b] \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \end{bmatrix} \quad (8)$$

Y 和二元随机变量 (X_1, X_2) 方差、协方差存在如下关系：

$$\text{var}(Y) = \text{var}(aX_1 + bX_2) = a^2 \text{var}(X_1) + b^2 \text{var}(X_2) + 2ab \text{cov}(X_1, X_2) \quad (9)$$

(9) 可以写成：

$$\text{var}(Y) = [a \ b] \underbrace{\begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \text{var}(X_2) \end{bmatrix}}_{\Sigma} \begin{bmatrix} a \\ b \end{bmatrix} \quad (10)$$

相信大家已经在上式中看到了如下协方差矩阵：

$$\Sigma = \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \text{var}(X_2) \end{bmatrix} \quad (11)$$

也就是说，(10) 可以写成：

$$\text{var}(Y) = [a \ b] \Sigma \begin{bmatrix} a \\ b \end{bmatrix} \quad (12)$$

D 维随机变量：朝单一方向投影

如果随机向量 $\chi = [X_1, X_2, \dots, X_D]^T$ 服从 $N(\mu_\chi, \Sigma_\chi)$, χ 在单位向量 v 方向上投影得到 Y :

$$Y = v^T \chi = v^T \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix} \quad (13)$$

Y 的期望 $E(Y)$ 为：

$$E(Y) = v^T \mu_\chi = v^T \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_D) \end{bmatrix} \quad (14)$$

Y 的方差 $E(Y)$ 为：

$$\text{var}(Y) = v^T \Sigma_\chi v \quad (15)$$

D 维随机变量：朝正交系投影

$\chi = [X_1, X_2, \dots, X_D]^T$ 服从 $N(\mu_\chi, \Sigma_\chi)$, χ 在规范正交系 V 投影得到 $\gamma = [Y_1, Y_2, \dots, Y_D]^T$:

$$\gamma = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_D \end{bmatrix} = V^T \chi = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_D^T \end{bmatrix} \chi = \begin{bmatrix} v_1^T \chi \\ v_2^T \chi \\ \vdots \\ v_D^T \chi \end{bmatrix} \quad (16)$$

γ 的期望 (质心) $E(\gamma)$ 为：

$$E(\gamma) = V^T \mu_\chi = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_D^T \end{bmatrix} \mu_\chi = \begin{bmatrix} v_1^T \mu_\chi \\ v_2^T \mu_\chi \\ \vdots \\ v_D^T \mu_\chi \end{bmatrix} \quad (17)$$

γ 的协方差矩阵 $\text{var}(\gamma)$ 为：

$$\text{var}(\gamma) = V^T \Sigma_{\chi} V = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix} \Sigma_{\chi} [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_D] = \begin{bmatrix} \mathbf{v}_1^T \Sigma_{\chi} \mathbf{v}_1 & \mathbf{v}_1^T \Sigma_{\chi} \mathbf{v}_2 & \cdots & \mathbf{v}_1^T \Sigma_{\chi} \mathbf{v}_D \\ \mathbf{v}_2^T \Sigma_{\chi} \mathbf{v}_1 & \mathbf{v}_2^T \Sigma_{\chi} \mathbf{v}_2 & \cdots & \mathbf{v}_2^T \Sigma_{\chi} \mathbf{v}_D \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_D^T \Sigma_{\chi} \mathbf{v}_1 & \mathbf{v}_D^T \Sigma_{\chi} \mathbf{v}_2 & \cdots & \mathbf{v}_D^T \Sigma_{\chi} \mathbf{v}_D \end{bmatrix} \quad (18)$$

上式还告诉我们， $\mathbf{v}_i^T \chi$ 和 $\mathbf{v}_j^T \chi$ 的协方差为：

$$\text{cov}(\mathbf{v}_i^T \chi, \mathbf{v}_j^T \chi) = \mathbf{v}_i^T \Sigma_{\chi} \mathbf{v}_j \quad (19)$$

14.3 单方向投影：鸢尾花两特征为例

本节以鸢尾花数据花萼长度、花萼宽度两特征为例讲解线性变换。我们首先看两个最简单的例子，将数据分别投影到横轴、纵轴。然后再看更一般的情况。

投影到 x 轴

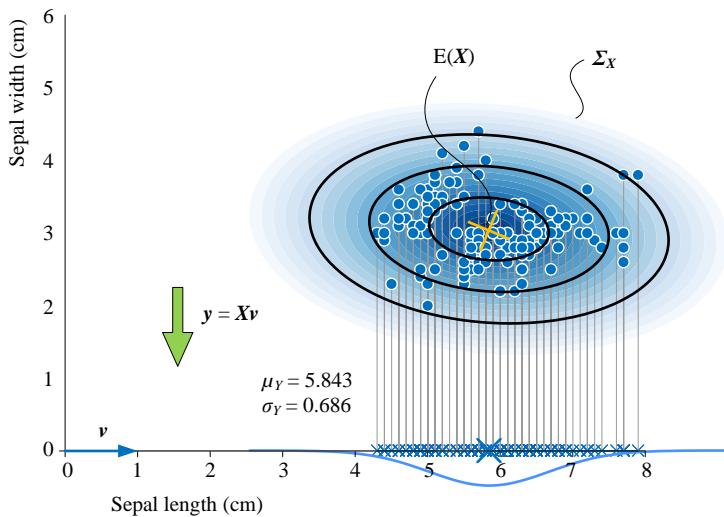
鸢尾花数据矩阵为 $X = [\mathbf{x}_1, \mathbf{x}_2]$ ，对应随机变量为 $\chi = [X_1, X_2]^T$ 。

如图 2 所示，将 X 投影到横轴，即：

$$\mathbf{y} = X\mathbf{v} = X \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \mathbf{x}_1 \quad (20)$$

从随机变量角度来看上述运算，即：

$$Y = \mathbf{v}^T \chi = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = X_1 \quad (21)$$

图 2. 逆时针 0 度, \mathbf{X} 向 \mathbf{v} 投影

\mathbf{X} 的质心为：

$$\mathbb{E}(\mathbf{X}) = [5.8433 \quad 3.0573] \quad (22)$$

由此计算得到图 2 中 \mathbf{y} 的质心为：

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{X})\mathbf{v} = [5.8433 \quad 3.0573] \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 5.8433 \quad (23)$$

\mathbf{X} 的协方差矩阵为：

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \quad (24)$$

由此计算得到图 2 中 \mathbf{y} 的方差为：

$$\text{var}(\mathbf{y}) = \mathbf{v}^T \text{var}(\mathbf{X}) \mathbf{v} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}^T \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 0.6856 \quad (25)$$

注意，图 2 中椭圆代表马氏距离。三个黑色旋转椭圆分别代表马氏距离为 1、2、3。

将图 2 中三个椭圆也投影到横轴上，大家会发现得到的三条线段分别代表 $\mu_1 \pm \sigma_1$ 、 $\mu_1 \pm 2\sigma_1$ 、 $\mu_1 \pm 3\sigma_1$ 。这绝不是几何上的巧合，本章后续会展开讲解。

投影到 \mathbf{y} 轴

如图 3 所示，将 \mathbf{X} 投影到纵轴，即：

$$\mathbf{y} = \mathbf{X}\mathbf{v} = [\mathbf{x}_1 \quad \mathbf{x}_2] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \mathbf{x}_2 \quad (26)$$

从随机变量角度来看上述运算，即：

$$Y = \mathbf{v}^T \boldsymbol{\chi} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = X_2 \quad (27)$$

计算图 3 中 \mathbf{y} 的质心为：

$$\mathbb{E}(\mathbf{y}) = \mathbb{E}(\mathbf{X})\mathbf{v} = [5.8433 \quad 3.0573] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 3.0573 \quad (28)$$

计算得到图 3 中 \mathbf{y} 的方差为：

$$\text{var}(\mathbf{y}) = \mathbf{v}^T \text{var}(\mathbf{X})\mathbf{v} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 0.1899 \quad (29)$$

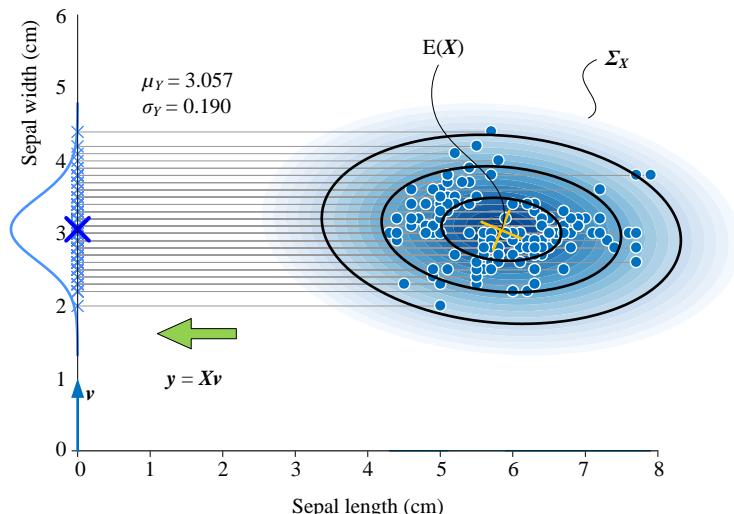


图 3. 逆时针 90 度, \mathbf{X} 向 \mathbf{v} 投影

其他情况

图 4 ~ 图 7 所示为其他四个投影场景，请大家自己分析。

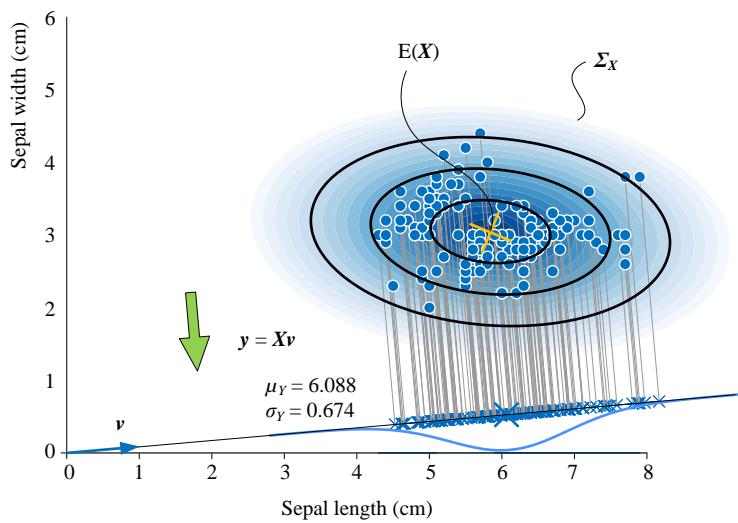


图 4. 逆时针 5 度, X 向 v 投影

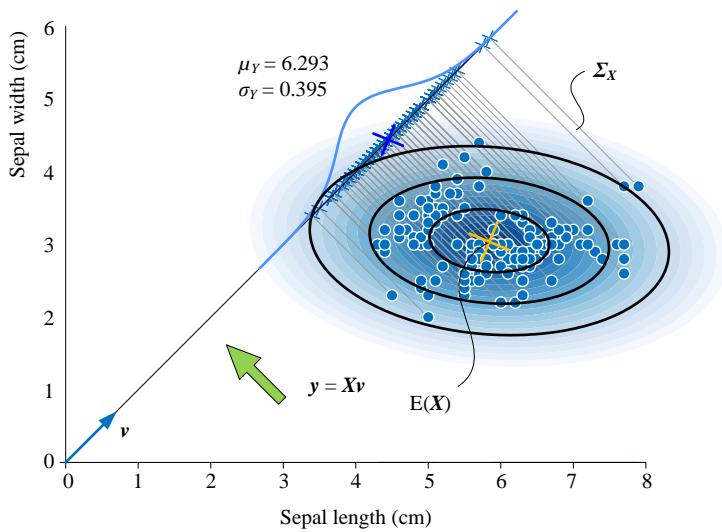


图 5. 逆时针 45 度, X 向 v 投影

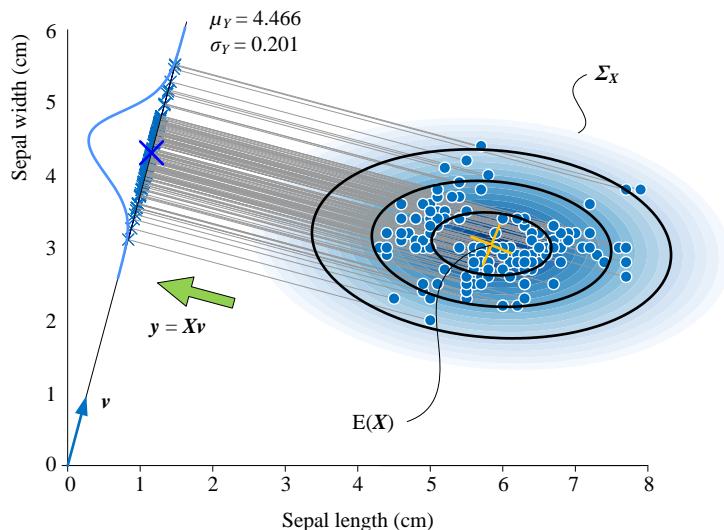


图 6. 逆时针 75 度, X 向 v 投影

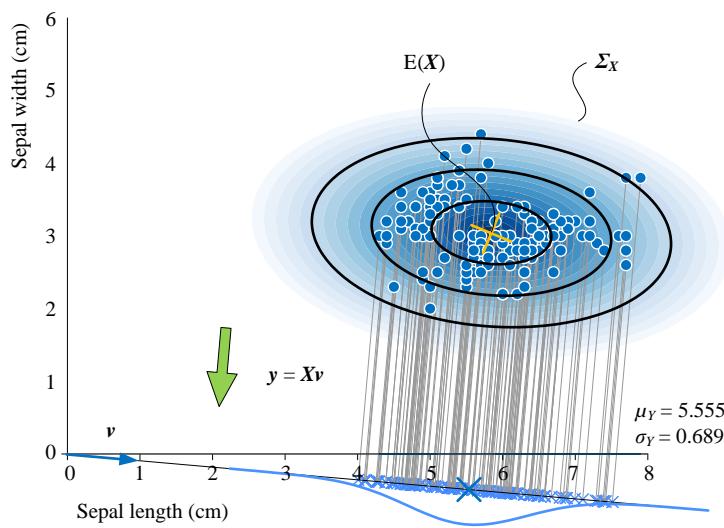


图 7. 逆时针 -5 度, X 向 v 投影



代码 Bk5_Ch14_01.py 绘制图 2 ~ 图 7。

14.4 正交系投影：鸢尾花两特征为例

正交系

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

给定正交系 V :

$$V = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \quad (30)$$

如图 8 所示，数据 X 可以投影到正交系 V 中得到数据 Y :

$$Y = X V \quad (31)$$

展开上式得到:

$$[y_1 \ y_2] = X [\nu_1 \ \nu_2] = [X\nu_1 \ X\nu_2] \quad (32)$$

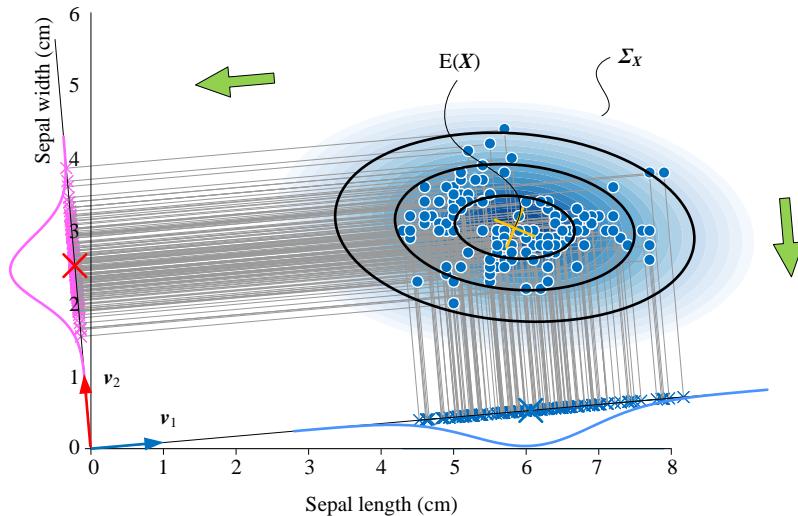


图 8. X 向正交系 V 投影

随机变量为 $\chi = [X_1, X_2]^T$ 投影到 V 得到 $\gamma = [Y_1, Y_2]^T$:

$$\gamma = V^T \chi \quad (33)$$

展开上式得到:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = V^T \chi = \begin{bmatrix} \nu_1^T \chi \\ \nu_2^T \chi \end{bmatrix} \quad (34)$$

注意比较 (31) 和 (33) 的转置关系。

向第一方向投影

先考虑 X 向 ν_1 投影:

$$\mathbf{v}_1 = \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \quad (35)$$

将数据 X 投影到 \mathbf{v}_1 得到：

$$\mathbf{y}_1 = \mathbf{X}\mathbf{v}_1 \quad (36)$$

类似地，将 $\chi = [X_1, X_2]^T$ 投影到 \mathbf{v}_1 得到 Y_1 ：

$$Y_1 = [X_1 \quad X_2] \mathbf{v}_1 = [X_1 \quad X_2] \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} = \cos \theta X_1 + \sin \theta X_2 \quad (37)$$

Y_1 的质心为：

$$\begin{aligned} E(Y_1) &= E(\mathbf{X})\mathbf{v}_1 = [5.8433 \quad 3.0573] \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \\ &\approx 3.0573 \times \sin(\theta) + 5.8433 \times \cos(\theta) \end{aligned} \quad (38)$$

Y_1 的方差为：

$$\begin{aligned} \text{var}(Y_1) &= \mathbf{v}_1^T \boldsymbol{\Sigma}_X \mathbf{v}_1 = [\cos \theta \quad \sin \theta] \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} \cos \theta \\ \sin \theta \end{bmatrix} \\ &\approx -0.0424 \times \sin(2\theta) + 0.2478 \times \cos(2\theta) + 0.4378 \end{aligned} \quad (39)$$

向第二方向投影

同理，给定 \mathbf{v}_2 ：

$$\mathbf{v}_2 = \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \quad (40)$$

将数据 X 投影到 \mathbf{v}_2 得到：

$$\mathbf{y}_2 = \mathbf{X}\mathbf{v}_2 \quad (41)$$

将 $\chi = [X_1, X_2]^T$ 投影到 \mathbf{v}_2 得到 Y_2 ：

$$Y_2 = [X_1 \quad X_2] \mathbf{v}_2 = [X_1 \quad X_2] \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} = -\sin \theta X_1 + \cos \theta X_2 \quad (42)$$

Y_2 的质心为：

$$\begin{aligned} \mu_{Y_2} &= E(\mathbf{X})\mathbf{v}_2 = [5.8433 \quad 3.0573] \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \\ &\approx -5.8433 \times \sin(\theta) + 3.0573 \times \cos(\theta) \end{aligned} \quad (43)$$

Y_2 的方差为：

$$\begin{aligned}\text{var}(Y_2) &= \mathbf{v}_2^T \boldsymbol{\Sigma}_X \mathbf{v}_2 = [-\sin \theta \quad \cos \theta] \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \\ &\approx 0.0424 \times \sin(2\theta) - 0.2478 \times \cos(2\theta) + 0.4378\end{aligned}\quad (44)$$

协方差

Y_1 和 Y_2 的协方差为：

$$\begin{aligned}\text{cov}(Y_1, Y_2) &= \mathbf{v}_1^T \boldsymbol{\Sigma}_X \mathbf{v}_2 = [\cos \theta \quad \sin \theta] \begin{bmatrix} 0.6856 & -0.0424 \\ -0.0424 & 0.1899 \end{bmatrix} \begin{bmatrix} -\sin \theta \\ \cos \theta \end{bmatrix} \\ &\approx -0.2478 \times \sin(2\theta) - 0.0424 \times \cos(2\theta)\end{aligned}\quad (45)$$

利用如下三角函数关系：

$$\begin{aligned}f(\theta) &= a \sin(\theta) + b \cos(\theta) \\ &= \sqrt{a^2 + b^2} \left(\frac{a}{\sqrt{a^2 + b^2}} \sin(\theta) + \frac{b}{\sqrt{a^2 + b^2}} \cos(\theta) \right) \\ &= \sqrt{a^2 + b^2} (\sin(\theta) \cos(\phi) + \cos(\theta) \sin(\phi)) \\ &= A \sin(\theta + \phi)\end{aligned}\quad (46)$$

其中，

$$\begin{aligned}\phi &= \arctan\left(\frac{b}{a}\right) \\ A &= \sqrt{a^2 + b^2}\end{aligned}\quad (47)$$

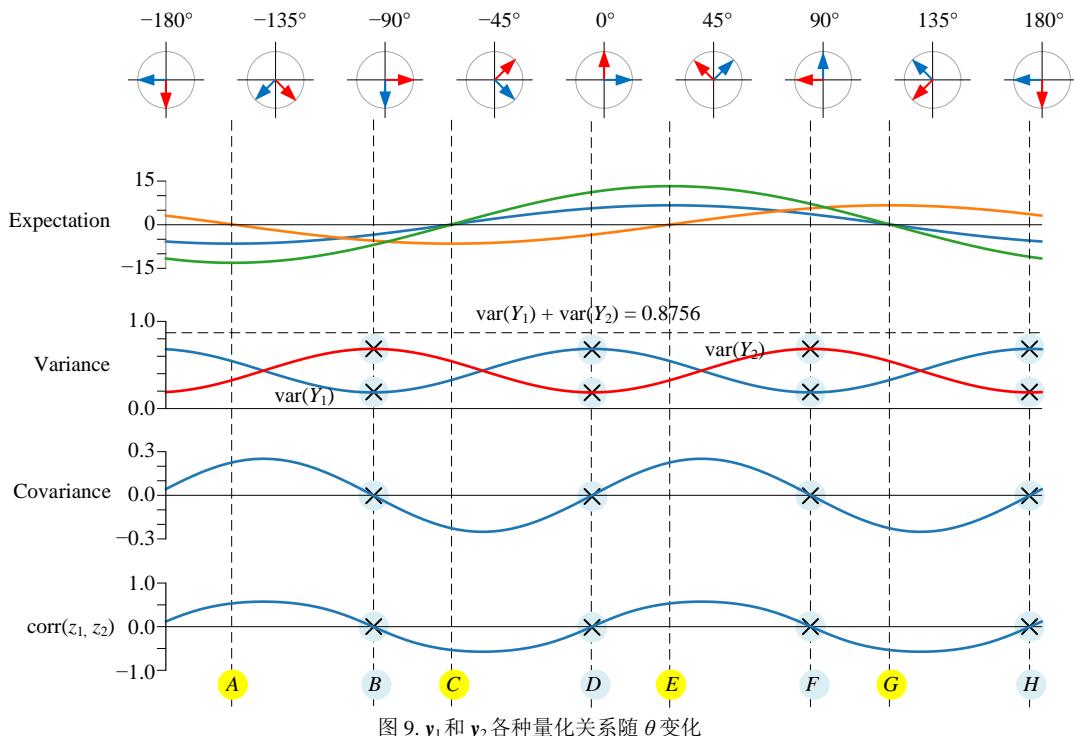
我们可以进一步整理 (38)、(39)、(43)、(44)、(45)，这部分推导交给大家完成。

如图 9 所示，期望、方差、协方差随 θ 变化。请大家特别注意 Y_1 和 Y_2 的方差之和为定值，即：

$$\begin{aligned}\text{var}(Y_1) + \text{var}(Y_2) &\approx -0.0424 \times \sin(2\theta) + 0.2478 \times \cos(2\theta) + 0.4378 + \\ &\quad 0.0424 \times \sin(2\theta) - 0.2478 \times \cos(2\theta) + 0.4378 \\ &\approx 0.8756\end{aligned}\quad (48)$$



以上内容实际上解释了鸢尾花书《矩阵力量》第 18 章中看到的曲线趋势。

图 9. y_1 和 y_2 各种量化关系随 θ 变化

协方差矩阵

$\gamma = [Y_1, Y_2]^T$ 的协方差矩阵 Σ_γ 为：

$$\text{var}(\gamma) = \Sigma_\gamma = V^T \Sigma_X V \quad (49)$$

图 10 所示为当 θ 取不同值时，协方差矩阵 Σ_γ 的三种不同可视化方案的变化情况。

特别地，如图 10 (b) 所示，当 θ 约为 -4.85 度时，协方差矩阵 Σ_γ 为对角方阵。这意味着 Y_1 和 Y_2 的相关性系数为 0。

在图 9 中，我们可以发现，当 θ 约为 -4.85 度时， $\text{var}(Y_1)$ 取得最大值， $\text{var}(Y_2)$ 取得最小值。如图 11 所示为数据矩阵在这个正交坐标系中投影的结果。这一点对于本章后续要讲解的主成分分析非常重要。

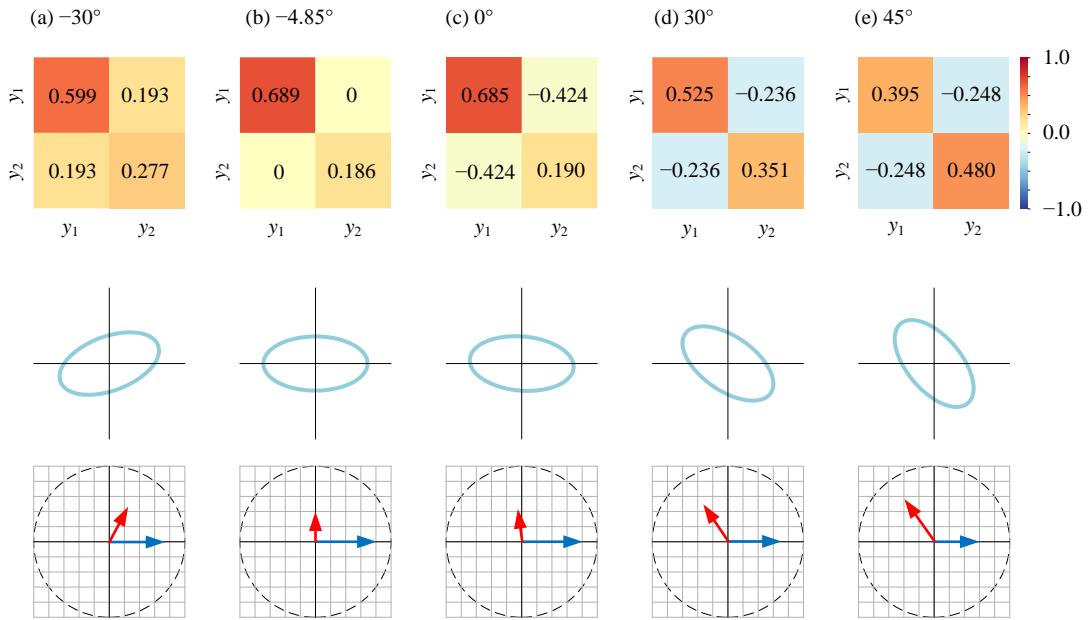
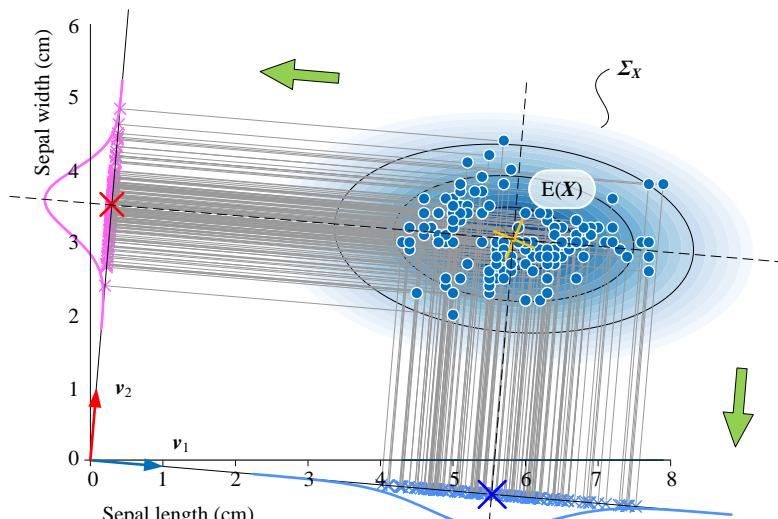


图 10. 协方差矩阵的可视化

图 11. X 向正交系 V 投影, -4.8575 度

14.5 以椭圆投影为视角看线性变换

本节将从椭圆投影视角理解随机变量的线性转换。

“正” 矩形

如图 12 所示，三个“正”矩形的四条边分别和马氏距离为 1、2、3 的椭圆相切。其中，和马氏距离为 1 的矩形相切的矩形的长、宽分别为 $2\sigma_1$ 、 $2\sigma_2$ 。

上一章提到过，图 12 中这个大矩形的面积为 $4\sigma_1\sigma_2$ ，其对角线长度为矩形对角线长度为 $2\sqrt{\sigma_1^2 + \sigma_2^2}$ 。

上一章特别强调图 12 中阴影区域对应的 1/4 矩形。这个 1/4 矩形的面积为 $\sigma_1\sigma_2$ ，1/4 矩形对角线长度为 $\sqrt{\sigma_1^2 + \sigma_2^2}$ ，这个值是其协方差矩阵的平方根 $\sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{\text{tr}(\Sigma_{2 \times 2})}$ 。

根据本章有关随机变量线性变换内容，如图 12 所示，这三个矩形“长边”所在位置分别对应 $\mu_1 \pm \sigma_1$ 、 $\mu_1 \pm 2\sigma_1$ 、 $\mu_1 \pm 3\sigma_1$ 。“宽边”所在位置分别对应 $\mu_2 \pm \sigma_2$ 、 $\mu_2 \pm 2\sigma_2$ 、 $\mu_2 \pm 3\sigma_2$ 。这并不是巧合，本节后续将用数学工具加以证明。

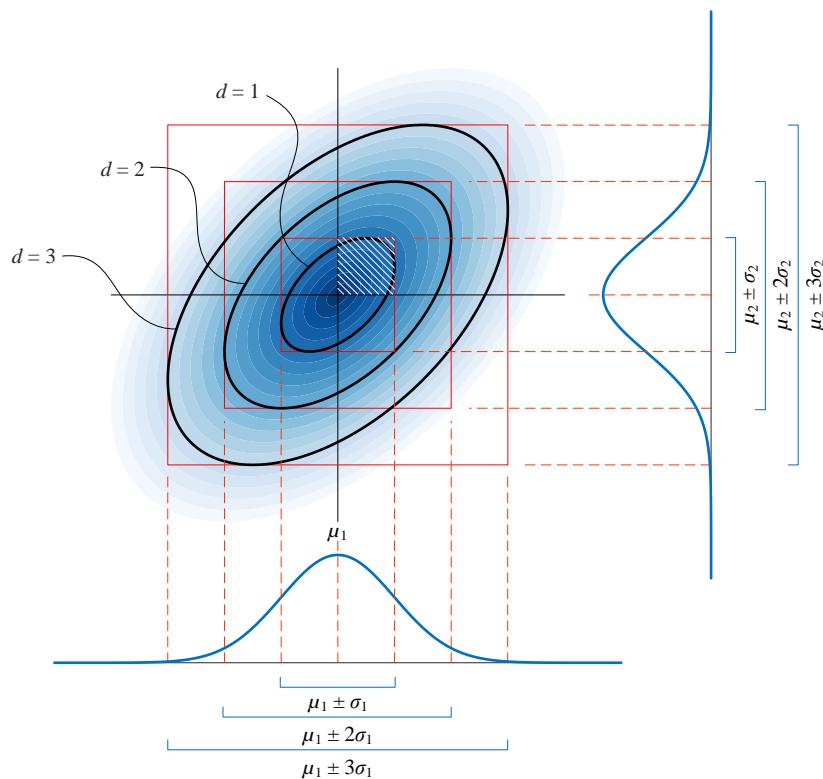


图 12. 和马氏距离椭圆相切的“正”矩形

“主轴” 矩形

如图 13 所示，和马氏距离为 1 的椭圆相切的矩形有无数个。观察这些矩形，大家能够发现它们的顶点位于正圆之上。这意味着这些矩形的对角线长度相同，都是 $2\sqrt{\sigma_1^2 + \sigma_2^2}$ 。想要证明这个观察，需要用到矩阵迹的性质，证明留给大家自行完成。

除了图 12 中的“正”矩形之外，还有一个“旋转”矩形特别值得我们关注。这就是图 14 所示的“主轴”矩形。之所以叫“主轴”矩形，是因为这个矩形的四条边平行于椭圆的两条主轴（长轴、短轴）。

而特征值分解协方差矩阵就是获得椭圆主轴方向、长轴长度、短轴长度的数学工具。请大家根据上一章内容自行分析图 14 中和马氏距离为 1 椭圆相切的“主轴”矩形的几何特征。

请大家回忆协方差特征值分解得到的特征值和投影获得的两个分布的方差、标准差关系。

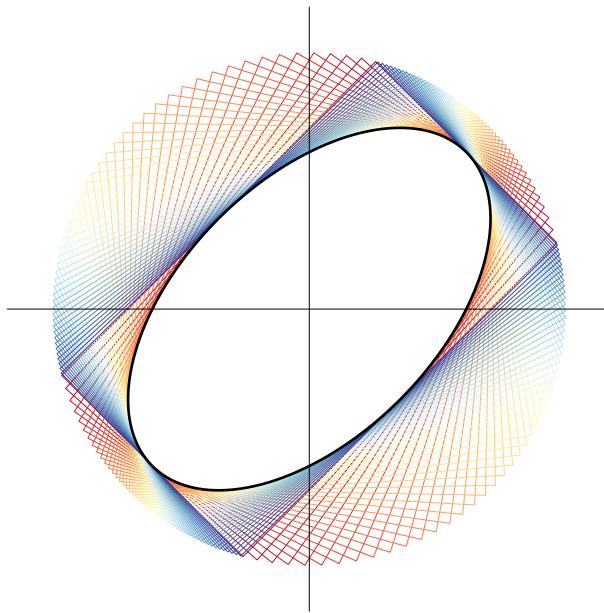


图 13. 和马氏距离椭圆相切的一组“旋转”矩形

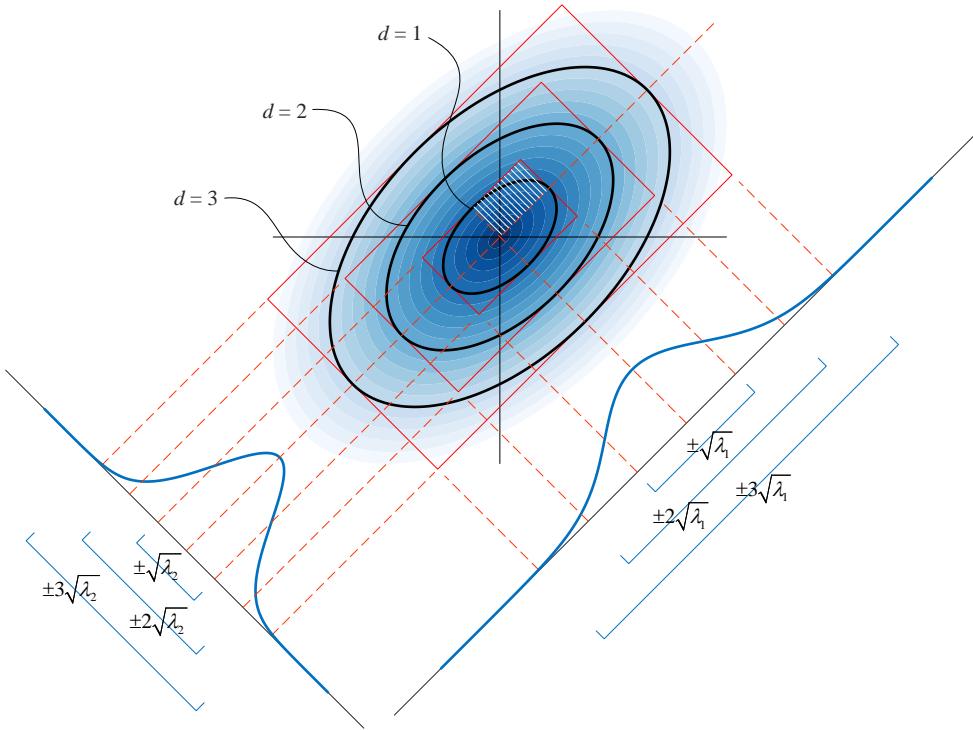


图 14. 和马氏距离椭圆相切的“主轴”矩形

椭圆切线

大家可能好奇如何绘制图 13 这组旋转矩形。如图 15 所示，首先，计算椭圆圆心 μ 和椭圆上任意一点 p 切线的距离 h 。 $2h$ 就是矩形一条边长度。

而切线的梯度向量 n 可以用来定位矩形的旋转角度。然后，根据矩形的对角线长度为 $2\sqrt{\sigma_1^2 + \sigma_2^2}$ ，我们便得到矩形另外一条边的长度。

问题来了，如何计算距离 h 和梯度向量 n ？

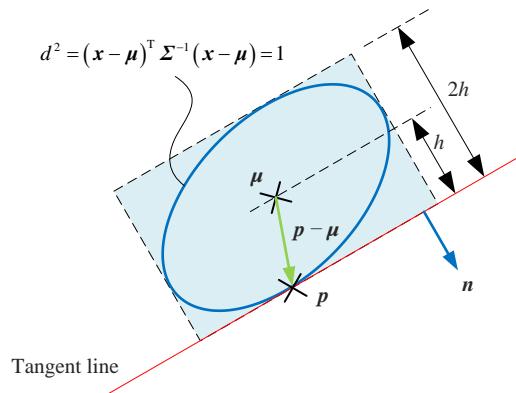


图 15. 计算马氏椭圆上任意一点切线原理



我们在《矩阵力量》第 20 章介绍过如何求解椭圆切线。

图 15 中椭圆的解析式为：

$$(x - \mu)^T \Sigma^{-1} (x - \mu) - 1 = 0 \quad (50)$$

p 在椭圆上，如下等式成立：

$$(p - \mu)^T \Sigma^{-1} (p - \mu) - 1 = 0 \quad (51)$$

定义如下函数 $f(x)$ ：

$$f(x) = (x - \mu)^T \Sigma^{-1} (x - \mu) - 1 = 0 \quad (52)$$

$f(x)$ 对 x 求偏导便得到梯度向量 n ：

$$n = \frac{\partial f(x)}{\partial x} = 2\Sigma^{-1}(x - \mu) \quad (53)$$

上式用到了《矩阵力量》第 17 章的多元微分。

也就是说，图 13 中椭圆上 p 点处切线的法向量为：

$$n = 2\Sigma^{-1}(p - \mu) \quad (54)$$

切点 p 和椭圆圆心 μ 的距离向量 $p - \mu$, 对应图 13 中的绿色箭头。而距离 h 就是向量 $p - \mu$ 在梯度向量 n 上的标量投影：

$$h = \frac{\mathbf{n}^T (\mathbf{p} - \boldsymbol{\mu})}{\|\mathbf{n}\|} \quad (55)$$

有了以上推导, 请大家自行编写代码绘制图 14。



《可视之美》一册详细讲解过这段可视化代码。

14.6 主成分分析：换个视角看数据

下面以鸢尾花数据作为原始数据, 从随机变量的线性变换角度理解主成分分析。

首先将鸢尾花萼长度、花萼宽度数据中心化, 即获得 $\mathbf{X}_c = \mathbf{X} - \mathbf{E}(\mathbf{X})$ 。图 16 所示为中心化数据的散点图。将数据投影到角度为逆时针 30 度的正交系 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2]$ 中。如前文所述, 数据投影到正交系中好比在 \mathbf{V} 中观察数据, 如图 17 所示。在 \mathbf{V} 中, 我们看到代表协方差矩阵的椭圆发生了明显旋转。在 \mathbf{v}_1 和 \mathbf{v}_2 方向上, 我们可以求得投影数据的分布情况。

图 18 ~ 图 23 所示为其他 3 组投影角度。请大家格外注意图 22 和图 23, 这就是前文说的最优化角度。这两幅图中的 \mathbf{v}_1 和 \mathbf{v}_2 分别为第一、第二主成分方向。

本章仅仅从随机变量的线性函数角度介绍主成分分析, 本书第 25 章将深入介绍主成分分析。



《矩阵力量》第 25 章介绍过, 特征值分解协方差矩阵仅仅是主成分分析六条基本技术路径之一, 《数据有道》还会介绍其他路径, 并做区分。

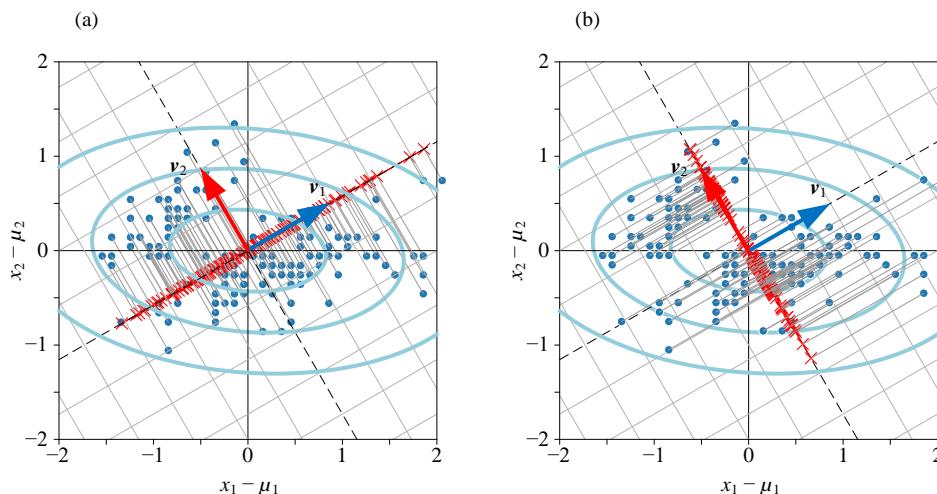


图 16. 正交系, 逆时针 30 度

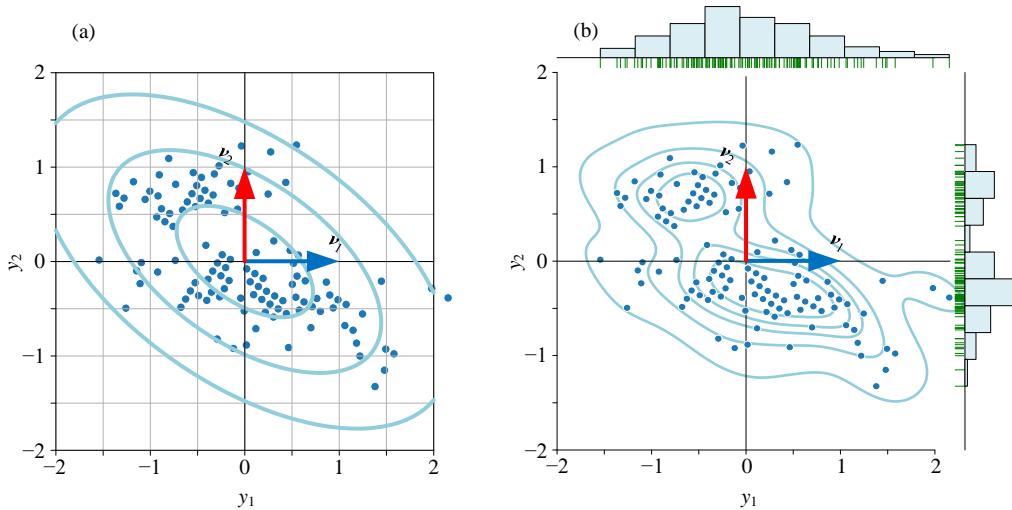


图 17. 数据顺时针旋转 30 度

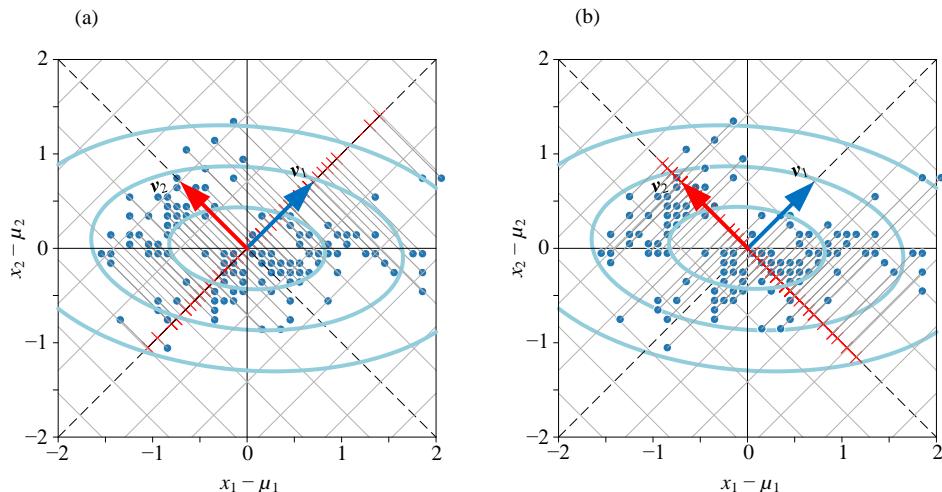


图 18. 正交系，逆时针 45 度

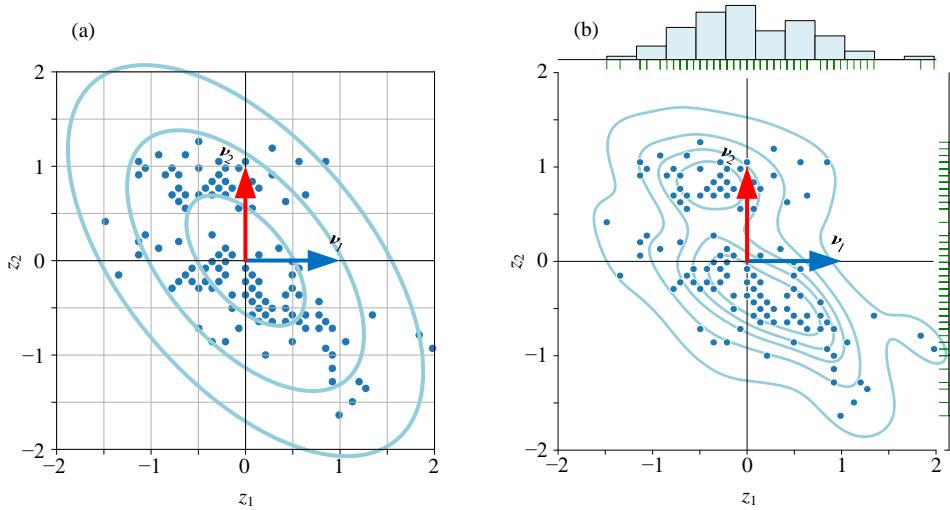


图 19. 数据顺时针旋转 45 度

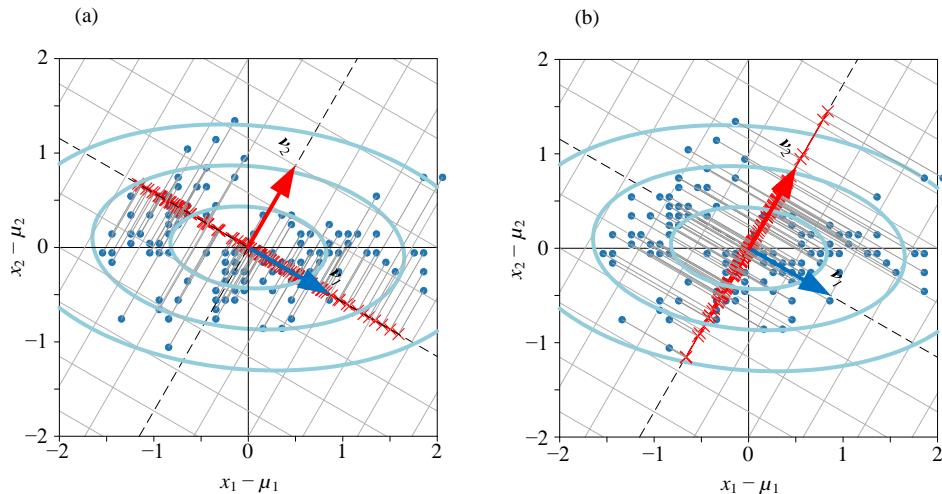


图 20. 正交系，逆时针-30 度

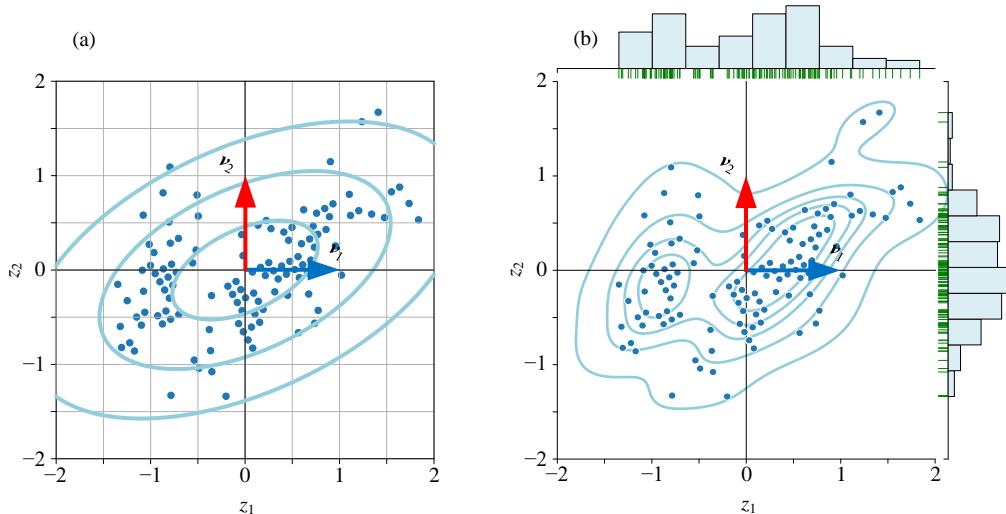


图 21. 数据顺时针旋转-30 度

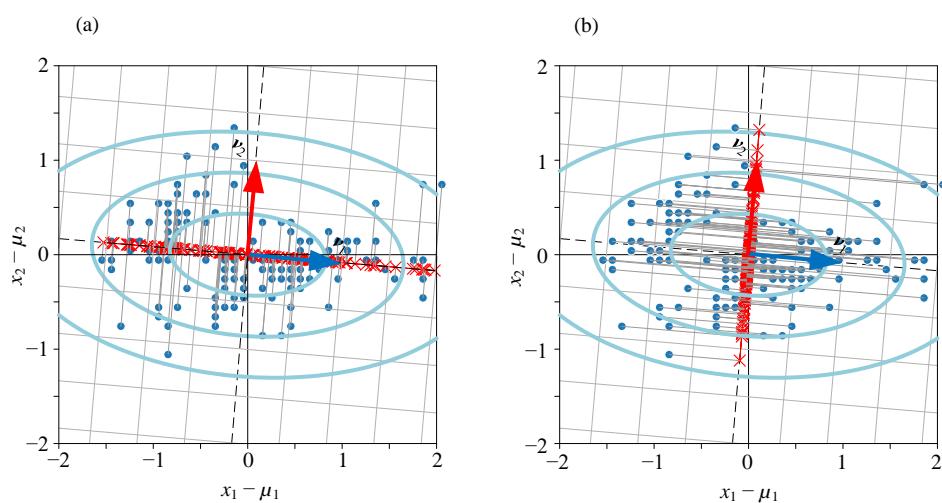


图 22. 正交系，逆时针旋转-4.85 度

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

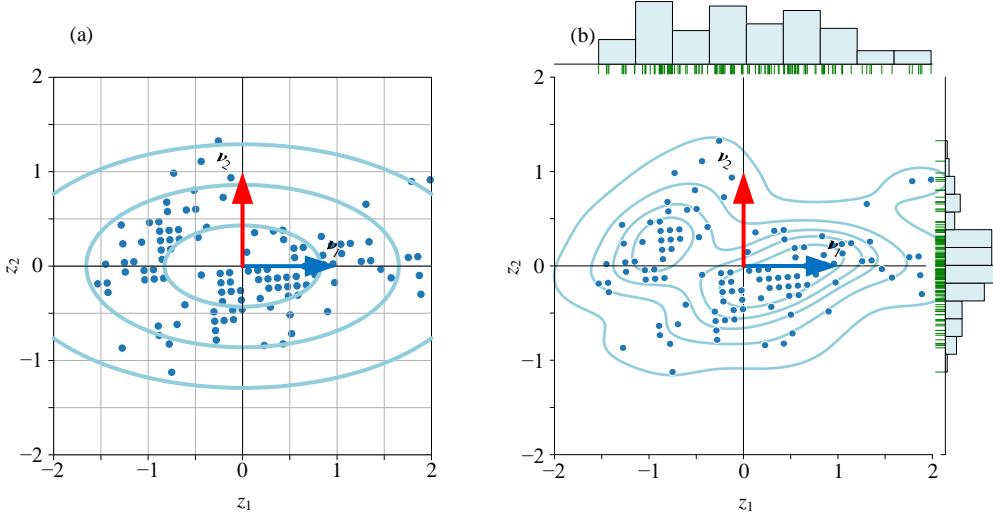
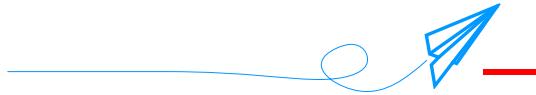


图 23. 数据顺时针旋转-4.85 度



随机变量的函数是指一个或多个随机变量组成的函数，其值也是一个随机变量。它们可以用来描述随机变量之间的关系或随机事件的性质。随机变量的函数在概率论和统计学中都有广泛的应用，例如用于建立概率模型、描述随机事件的分布和性质、进行概率推断和预测等。这一章，我们特别关注的是随机变量的线性变换。这是指将一个随机变量通过一个线性函数转化为另一个随机变量的过程，相当于线性代数中的仿射变换。

随机变量的线性变换在统计学和概率论中经常被用来描述随机变量之间的关系，例如线性回归模型、协方差矩阵和主成分分析等。通过线性变换，可以将随机变量从原始空间中转换到一个新的空间，从而发现不同随机变量之间的联系和规律。

请大家特别重视通过投影、椭圆视角理解随机变量的线性变换。

15

Monte Carlo Simulation

蒙特卡洛模拟

以概率统计为基础，基于伪随机数，进行数值模拟



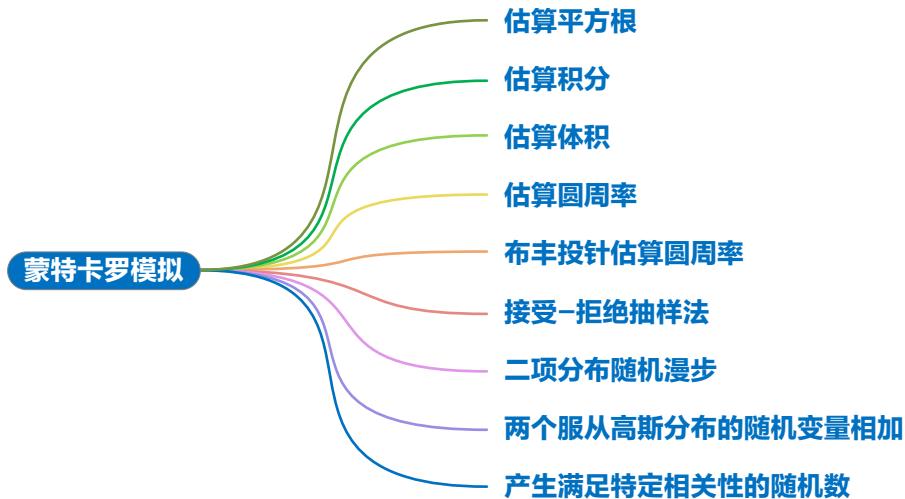
任何考虑用算术手段来产生随机数的人当然都是有原罪的。

Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.

—— 约翰·冯·诺伊曼 (John von Neumann) | 美国籍数学家 | 1903 ~ 1957



- ◀ matplotlib.patches.Circle() 绘制正圆
- ◀ matplotlib.pyplot.semilogx() 横轴设置为对数坐标
- ◀ numpy.empty() 产生全为 NaN 序列
- ◀ numpy.random.beta() 产生服从 Beta 分布的随机数
- ◀ numpy.random.binomial() 产生服从二项分布的随机数
- ◀ numpy.random.dirichlet() 产生服从 Dirichlet 分布的随机数
- ◀ numpy.random.exponential() 产生服从指数分布的随机数
- ◀ numpy.random.geometric() 产生服从几何分布的随机数
- ◀ numpy.random.lognormal() 产生服从对数正态分布的随机数
- ◀ numpy.random.multivariate_normal() 产生服从多项正态分布的随机数
- ◀ numpy.random.normal() 产生服从正态分布的随机数
- ◀ numpy.random.poisson() 产生服从泊松分布的随机数
- ◀ numpy.random.randint() 产生均匀整数随机数
- ◀ numpy.random.standard_t() 产生服从学生 t-分布的随机数
- ◀ numpy.random.uniform() 产生服从连续均匀分布的随机数
- ◀ numpy.where() 返回满足条件的元素序号
- ◀ scipy.integrate.dblquad() 求解双重定积分值
- ◀ scipy.integrate.quad() 求解定积分值
- ◀ scipy.linalg.cholesky() 对矩阵进行 Cholesky 分解
- ◀ seaborn.distplot() 绘制频率直方图和 KDE 曲线
- ◀ seaborn.heatmap() 绘制热图



15.1 蒙特卡洛模拟：基于伪随机数发生器

蒙特卡洛模拟 (Monte Carlo simulation)，也称统计模拟方法，是以概率统计理论为核心的数值计算方法。蒙特卡洛模拟将提供多种可能的结果以及通过大量随机数据样本得出的每种结果的概率。[冯·诺伊曼](#) (John von Neumann) 等三名科学家在 20 世纪 40 年代发明了蒙特卡洛模拟。他们以摩纳哥著名的赌城——[蒙特卡洛](#) (Monte Carlo)——为其命名。

[表 1](#) 所示为 NumPy 中和随机数有关的常见函数。

本章介绍几个最基本的蒙特卡洛模拟试验。

[表 1. NumPy 中和随机数有关的常见函数](#)

函数名称	函数介绍
numpy.random.beta()	生成指定形状参数的贝塔分布随机数
numpy.random.binomial()	返回给定形状的随机二项分布数组。
numpy.random.chisquare()	生成指定自由度的卡方分布随机数
numpy.random.choice()	随机从给定的数组中选择元素。
numpy.random.dirichlet()	生成指定参数的狄利克雷分布随机数
numpy.random.exponential()	生成指定尺度的指数分布随机数
numpy.random.gamma()	生成指定形状和尺度的伽马分布随机数
numpy.random.lognormal()	生成指定均值和标准差的对数正态分布随机数
numpy.random.multivariate_normal()	生成多元正态分布随机数
numpy.random.normal()	生成指定均值和标准差的正态分布随机数
numpy.random.poisson()	生成指定均值的泊松分布随机数
numpy.random.power()	返回给定形状的随机幂律分布数组。
numpy.random.rand()	返回一个给定形状的随机浮点数数组，值在 0 到 1 之间。
numpy.random.randint()	返回一个给定形状的随机整数数组，值在给定范围之间。
numpy.random.randn()	返回一个给定形状的随机浮点数数组，值遵循标准正态分布。
numpy.random.random()	生成 [0, 1] 之间的随机数
numpy.random.seed()	设置随机数生成器的种子，确保随机数生成的可重复性。
numpy.random.shuffle()	随机打乱给定的数组。
numpy.random.uniform()	生成指定范围内的均匀分布随机数

15.2 估算平方根

本节用蒙特卡洛模拟估算 $\sqrt{2}$ 。如图 1 所示，为了估算 $\sqrt{2}$ ，可以在 0 ~ 2 的范围内产生大量服从均匀分布的随机数。在 0 ~ 2 范围内，随机数在 $0 \sim \sqrt{2}$ 出现的概率为 $\sqrt{2}/2$ ， $\sqrt{2}$ 则可以根据下式估计得到：

$$\sqrt{2} \approx 2 \times \frac{n(0 \leq x \leq \sqrt{2})}{n(0 \leq x \leq 2)} \quad (1)$$

其中 $n()$ 计算频数。

由于 $\sqrt{2}$ 未知，所以采用图 1 所示平方技巧，即 $\sqrt{2}$ 可以根据下式得到：

$$\sqrt{2} \approx 2 \times \frac{n(0 \leq x^2 \leq 2)}{n(0 \leq x^2 \leq 4)} \quad (2)$$

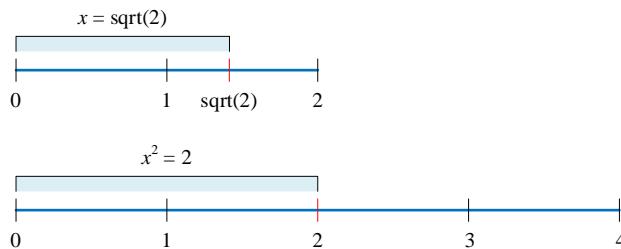
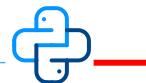


图 1. 估算 $\sqrt{2}$



代码文件 Bk5_Ch15_01.py 估算 $\sqrt{2}$ 。

15.3 估算积分

本节给出的例子用蒙特卡罗模拟方法估算积分。

给出如下函数 $f(x)$ ：

$$f(x) = \frac{x \cdot \sin(x)}{2} + 8 \quad (3)$$

计算 $f(x)$ 在 $[2, 10]$ 区间内定积分：

$$\int_2^{10} \left(\frac{x \cdot \sin(x)}{2} + 8 \right) dx \quad (4)$$

如图 2 所示，在 $[2, 10]$ 区间中，函数 $f(x)$ 的最大值为 12。在横轴取值从 2 ~ 10，纵轴取值从 0 ~ 12 的长方形空间里，产生满足均匀分布的 1000 个数据点。图 2 中蓝色 ● 在曲线之下，红色 ✕ 在曲线之上。图 2 中整个长方形的面积为 96，定积分对应曲线之下的面积 A ，可以通过下式估算得到：

$$A \approx 96 \times \frac{n(\text{below } f(x))}{1000} \quad (5)$$

$n(\text{below } f(x))$ 为 1000 个数据点中位于 $f(x)$ 曲线之下的数量。

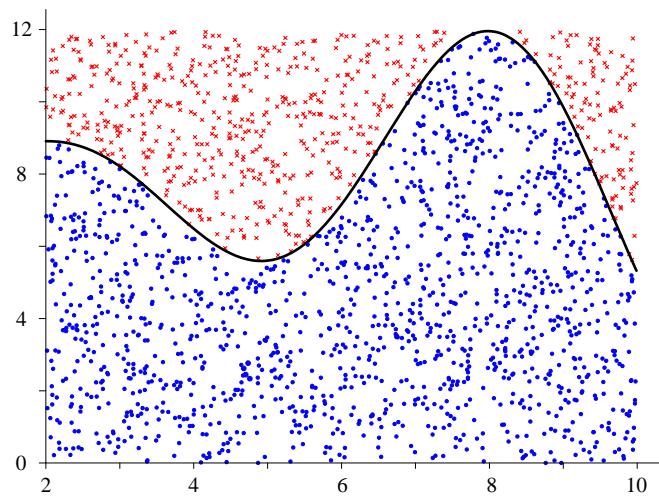


图 2. 用蒙特卡罗模拟法求积分



代码文件 Bk5_Ch15_02.py 估算积分。

15.4 估算体积

本节用蒙特卡洛模拟估算空间体积大小。图 3 (a) 所示二次曲面解析式如下：

$$z = 2 - x^2 - y^2 \quad (6)$$

当 x 和 y 均在 $[-1, 1]$ 范围内时，编写代码用蒙特卡洛模拟估算图 3 (a) 曲面和 $z = 0$ 平面（蓝色）构造的空间体积。这个体积相当于如下双重定积分：

$$\int_{-1}^1 \int_{-1}^1 (2 - x^2 - y^2) dx dy \quad (7)$$

整个立方体空间体积为 8，在这个空间均匀产生 5000 个随机点。如图 3 (b) 所示，二次曲面之上随机点为红色，曲面之下随机点为蓝色。类似上一节，根据随机点的比例，可以估算 (7) 定积分。

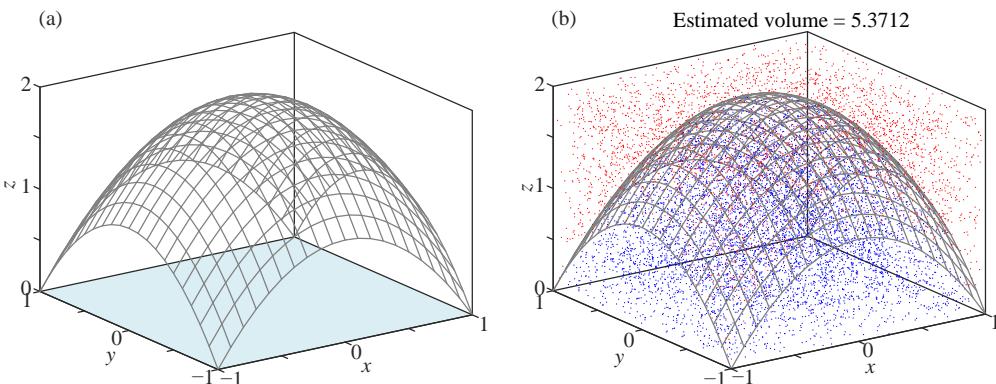


图 3. 利用蒙特卡洛估算体积



Bk5_Ch15_03.py 估算本节体积。

15.5 估算圆周率

本节介绍采用蒙特卡罗模拟法估算圆周率。



《数学要素》一本已经介绍几种方法估算圆周率 π ，请大家回忆。

圆面积和正方形面积之间的比例关系为：

$$\frac{A_{\text{circle}}}{A_{\text{square}}} = \frac{\pi}{4} \quad (8)$$

可以推导得到：

$$\pi = 4 \times \frac{A_{\text{circle}}}{A_{\text{square}}} \quad (9)$$

图 4 所示为一次随机数数量为 500 条件下，圆周率估算结果。图 5 所示为不断增大随机数数量，圆周率估算精确度不断提高。

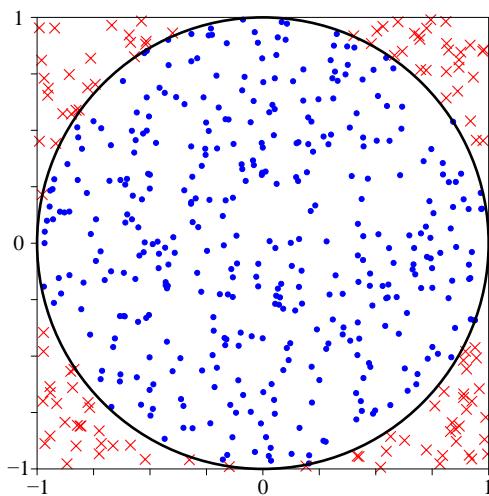


图 4. 蒙特卡罗模拟估算圆周率

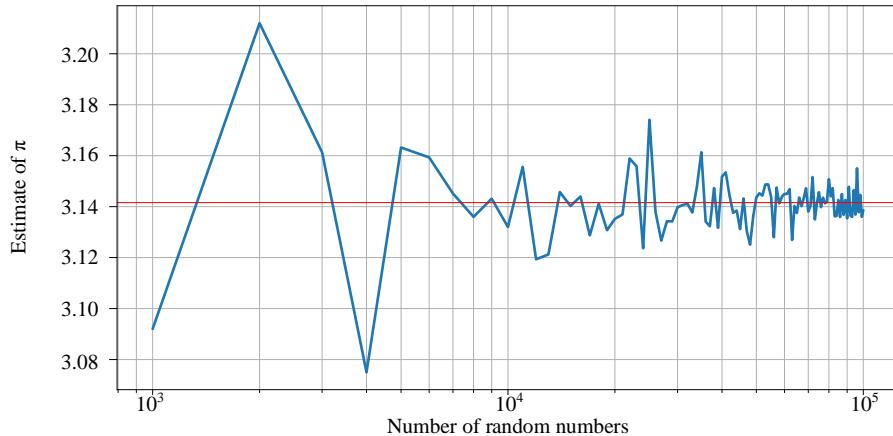
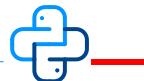


图 5. 不断增大随机数数量，圆周率估算精确度不断提高



Bk5_Ch15_04.py 利用蒙特卡洛模拟估算圆周率。

15.6 布丰投针估算圆周率

布丰投针 (Buffon's needle problem) 也可以用来估算圆周率。

十八世纪，法国博物学家**布丰** (Comte de Buffon) 提出著名的布丰投针问题。一个用平行且等距木纹铺成的地板，随意投掷一支长度比木纹间距略小的针，求针和其中一条木纹相交的概率。

如图6所示，和平行线相交的针颜色为红色，不和平行线相交的针颜色为蓝色。设平行线距离为 t ，针的长度为 l 。本节布丰投针问题，我们仅仅考虑“短针”情况，即 $l < t$ 。

如放大视图所示， x 为针的中心和最近平行线的距离， θ 为针和平行线之间的不大于 90° 夹角。

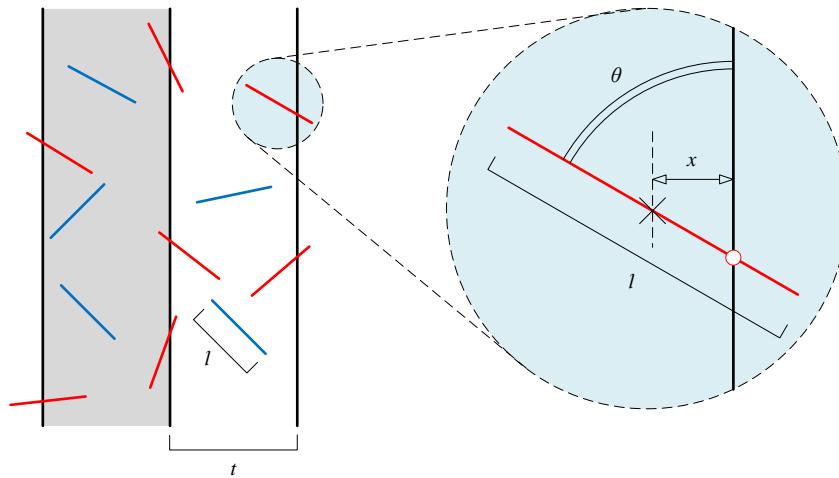


图 6. 布丰投针原理

不难理解， X 作为一个随机变量是在 $[0, t/2]$ 区间的连续均匀分布，概率密度函数为：

$$f_X(x) = \frac{2}{t} \quad x \in [0, t/2] \quad (10)$$

同理， Θ 作为一个随机变量是在 $[0, \pi/2]$ 区间的均匀分布，概率密度函数为：

$$f_\Theta(\theta) = \frac{2}{\pi} \quad \theta \in [0, \pi/2] \quad (11)$$

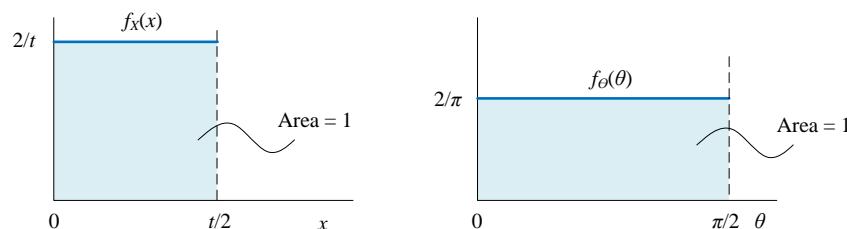


图 7. X 和 Θ 的概率密度函数

显然， X 和 Θ 这两个随机变量相互独立；因此，它们的联合概率密度函数是两者之积，即：

$$f_{\theta,x}(\theta, x) = \frac{2}{\pi} \frac{2}{t} = \frac{4}{\pi t} \quad \theta \in [0, \pi/2], \quad x \in [0, t/2] \quad (12)$$

给定夹角 θ , 满足如下条件, 针和平行线相交:

$$x \leq \frac{l}{2} \sin \theta, \quad \theta \in [0, \pi/2], \quad x \in [0, t/2] \quad (13)$$

因此, 针线相交的概率为如下双重定积分:

$$\Pr(\text{cross}) = \int_0^{\pi/2} \int_0^{l/\sin \theta} \frac{4}{\pi t} dx d\theta = \frac{2l}{\pi t} \quad (14)$$

假设抛 n 根针, 其中有 c 根和平行线相交, 概率值 $\Pr(\text{cross})$ 可以通过下式估算:

$$\Pr(\text{cross}) \approx \frac{c}{n} \quad (15)$$

联立 (14) 和 (15), 可以得到:

$$\frac{2l}{\pi t} \approx \frac{c}{n} \quad (16)$$

从而推导得到, 圆周率的估算值:

$$\pi \approx \frac{2l}{t} \frac{n}{c} \quad (17)$$

图 8 所示为某次试验投掷 2000 根针, 612 根和平行线相交 (红色线); 式样中, 针的长度 $l=1$, 平行线间隔 $t=2$ 。

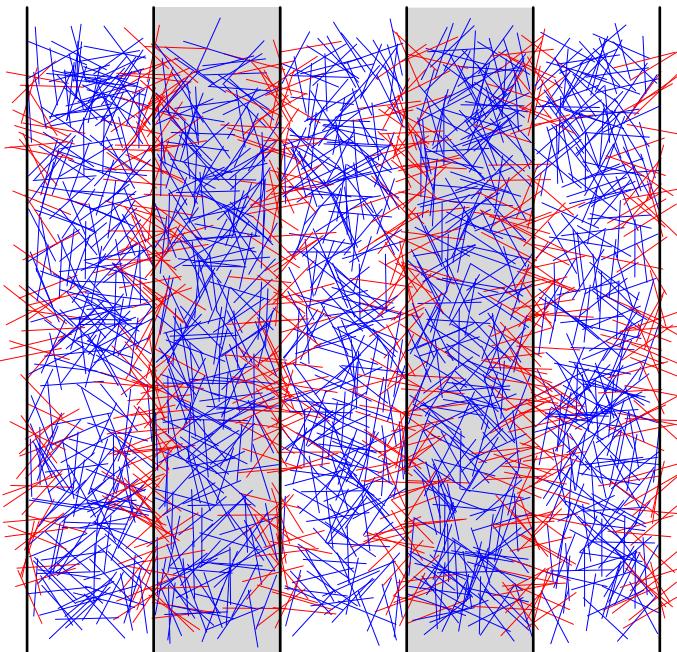


图 8. 投掷 2000 根针，612 根和平行线相交，针的长度 $l = 1$ ，平行线间隔 $t = 2$

实际上，根据(13)，我们知道针和平行线相交的概率 $\text{Pr}(\text{cross})$ 可以进一步简化。在图9阴影区域产生满足均匀分布的随机数，随机数落入蓝色区域的概率就是 $\text{Pr}(\text{cross})$ 。这样，我们可以根据这一思路编程解决这个简化版的布丰投针估算圆周率问题。

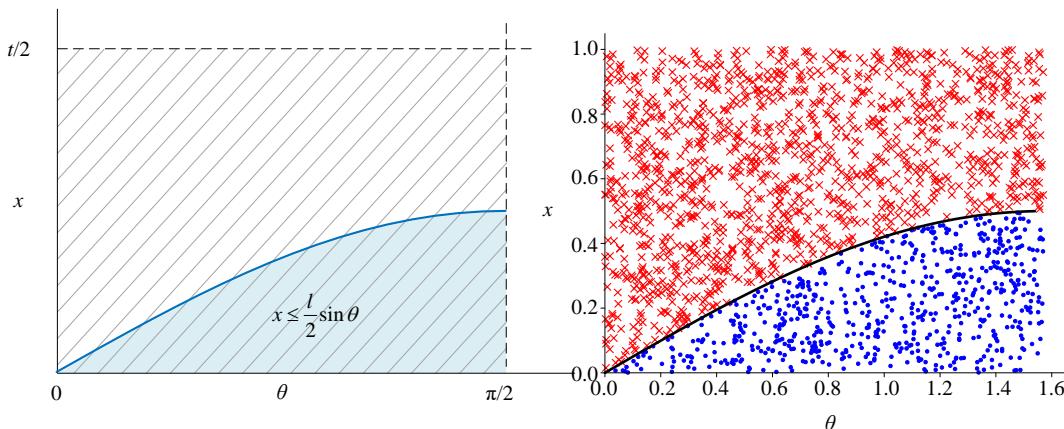


图 9. 针和平行线相交的概率，蒙特卡洛模拟试验结果



Bk5_Ch15_05.py 完成简化版布丰投针蒙特卡洛模拟试验。

15.7 接受-拒绝抽样法

随机变量 X 的概率密度函数为 $f_X(x)$ ，但是 $f_X(x)$ 不可以直接抽样。也就是说，不能直接产生满足 $f_X(x)$ 的随机数。我们可以采用本节介绍的**接受-拒绝抽样法** (accept-reject sampling method)。接受-拒绝抽样法适合于概率密度函数复杂、不能直接抽样的情况。

接受-拒绝抽样法的基本思想是，生成一个**辅助分布** (proposal distribution)，并利用这个分布来生成随机数。

然后，计算目标概率分布在该点处的概率密度，并将其除以辅助分布在该点处的概率密度，得到**接受率** (acceptance ratio)。

随机生成一个介于 0 和 1 之间的均匀分布的随机数，如果这个随机数小于接受率，则接受这个样本，否则拒绝。重复此过程，直到生成足够多的样本。

接受-拒绝抽样法的优点是简单易用、适用范围广，可以应用于各种不同的概率分布。它的缺点是样本生成的效率可能较低，因为需要进行接受和拒绝的判断。在实践中，辅助分布的选择对于样本生成的效率和精度非常重要。

给定如图 10 所示的随机变量 X 的概率密度函数为 $f_X(x)$ 。显然没有“现成”的随机数发生器能够直接生成满足 $f_X(x)$ 的随机数。

我们首先生成如图 11 所示的连续均匀分布随机数。简单来说，如图 12 所示，接受-拒绝抽样法就是在图 11 中“剪裁”得到形似图 10 的部分，并“接受”这些随机数。图 13 所示为用直方图可视化“拒绝”和“接受”部分的随机数。大家很容易发现，图中浅蓝色部分矩形构成形状形似图 10 中的 $f_X(x)$ 。



我们将在本书第 22 章用到接受-拒绝抽样法。

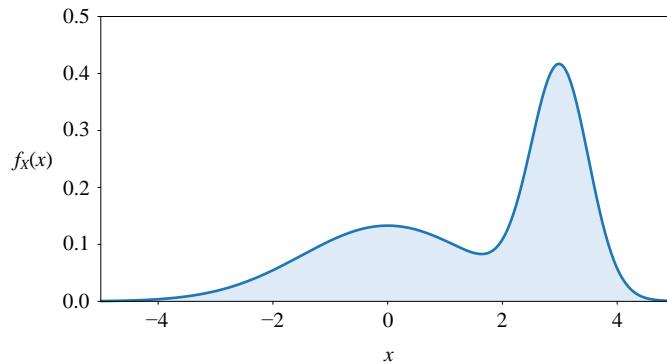


图 10. 随机变量 X 的概率密度函数

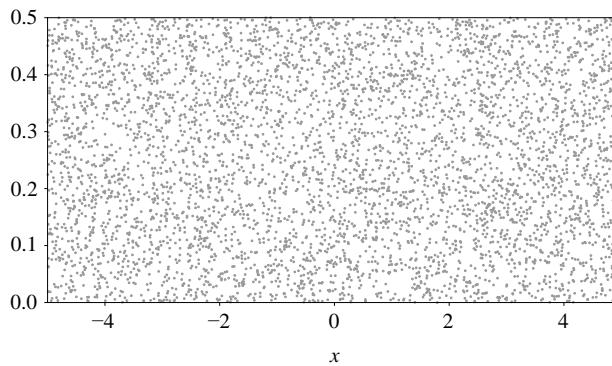


图 11. 生成连续均匀分布随机数

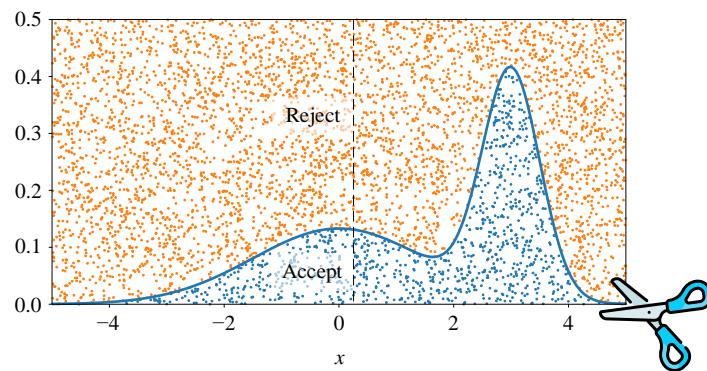


图 12. “剪裁”连续均匀分布随机数

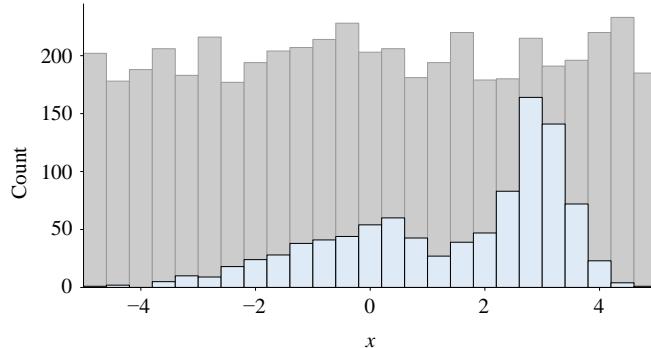


图 13. “接受” vs “拒绝”部分随机数

15.8 二项分布随机漫步



“鸢尾花书”《数学要素》第 20 章讲过在二叉树规定的网格行走的例子。

如图 14 所示，登山者在二叉树始点或中间节点时，他都会面临“向上”或“向下”抉择。如果登山者，通过抛硬币来决定每一步的行走路径——正面，向右上走；反面，向右下走。

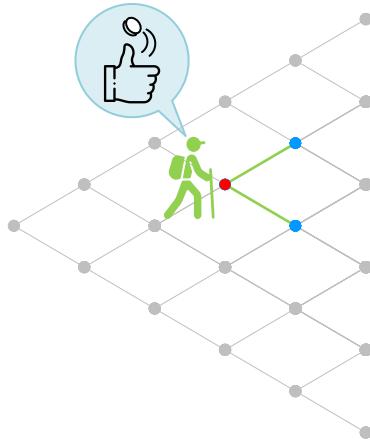
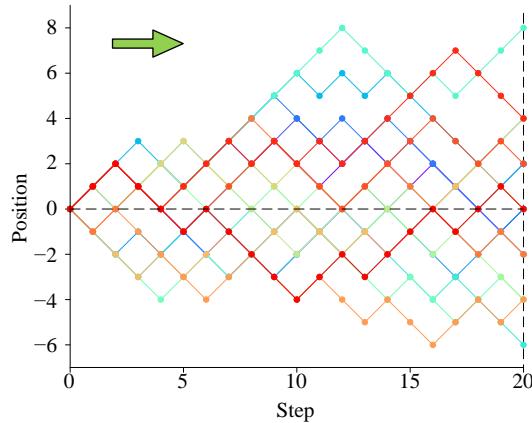
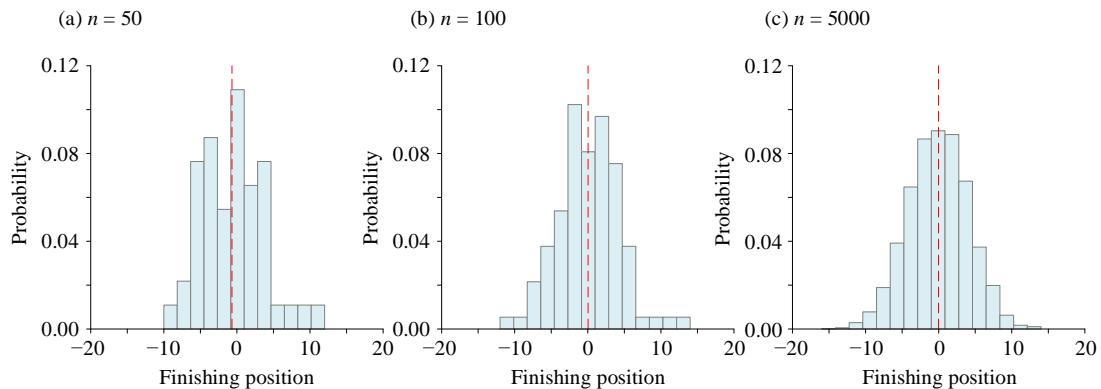
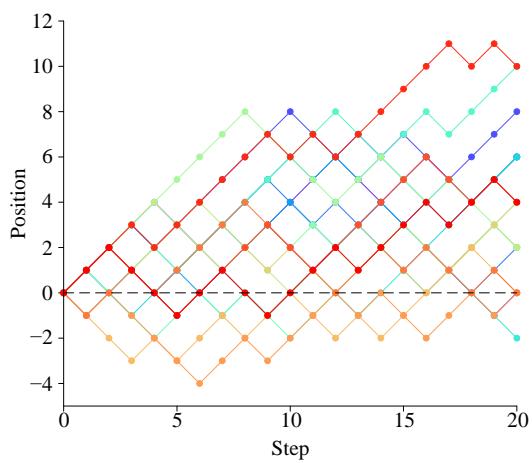


图 14. 二叉树路径与可能性，图片来自《数学要素》

图 15 所示为若干条二叉树随机行走路径，模拟时向上行走的概率 $p = 0.5$ 。乍一看图 15 很难发现任何规律。但是不断增大随机行走的路径数 n ，如图 16 所示，我们发现登山者到达终点的位置呈现类似二项分布规律。观察图 16 (c)，我们发现当 $p = 0.5$ 时，登山者大概率会到达二叉树网格终点中部。

图 17、图 18 对应登山者向上行走的概率 $p = 0.6$ 。图 19、图 20 对应登山者向上行走的概率 $p = 0.4$ 。请大家自行分析这四幅图。

图 15. 二叉树随机行走路径，向上行走的概率 $p = 0.5$ 图 16. 第 20 步时随机漫步位置分布， $p = 0.5$ 图 17. 二叉树随机行走路径，向上行走的概率 $p = 0.6$

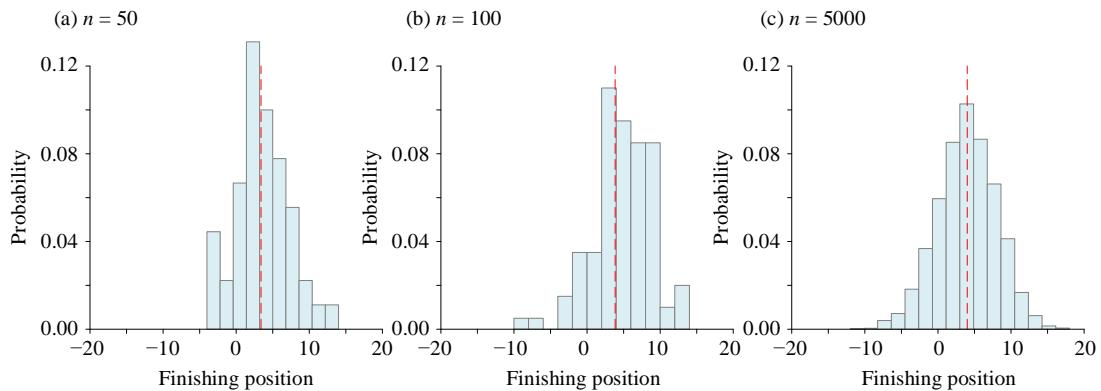
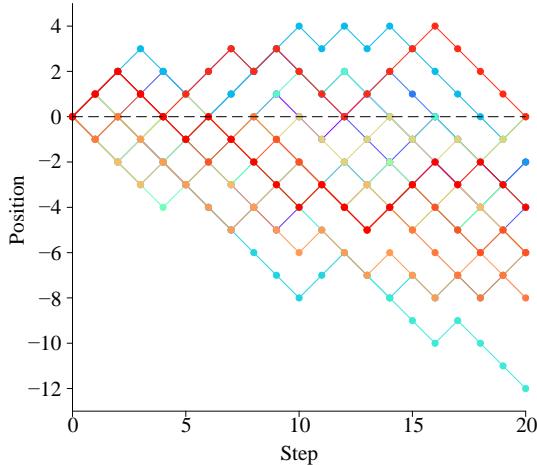
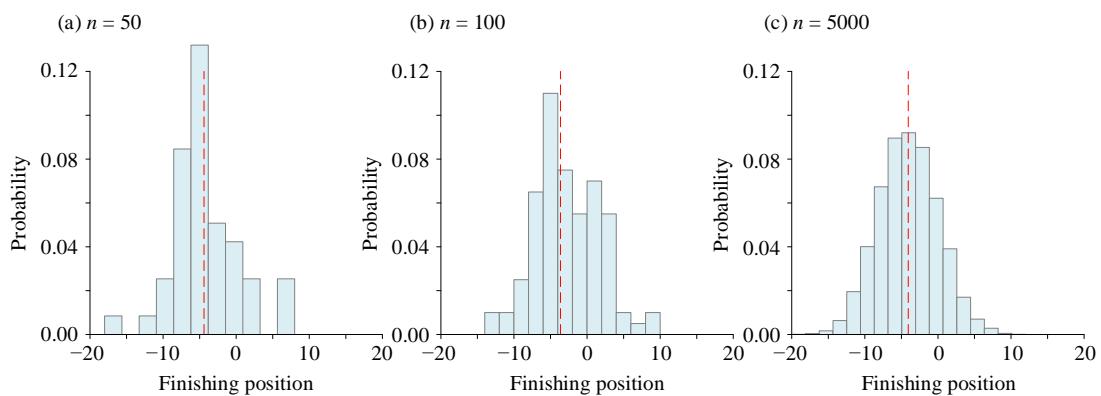
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

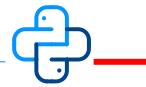
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 18. 第 20 步时随机漫步位置分布, $p = 0.6$ 图 19. 二叉树随机行走路径, 向上行走的概率 $p = 0.4$ 图 20. 向上行走的概率 $p = 0.4$



Bk5_Ch15_06.py 完成本节二叉树随机漫步试验。

15.9 两个服从高斯分布的随机变量相加

X_1 和 X_2 分别服从正态分布，具体如下：

$$\begin{cases} X_1 \sim N(\mu_1, \sigma_1^2) \\ X_2 \sim N(\mu_2, \sigma_2^2) \end{cases} \quad (18)$$

图 21 所示为 X_1 和 X_2 的随机数分布情况。

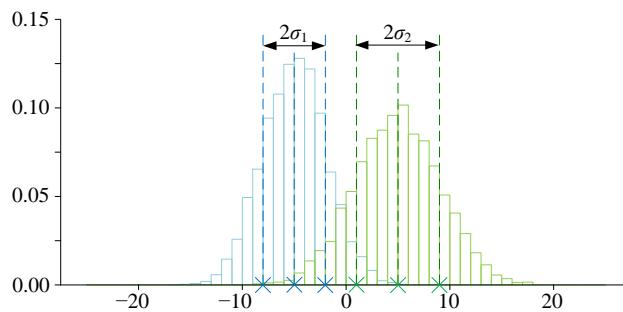


图 21. X_1 和 X_2 的随机数分布情况

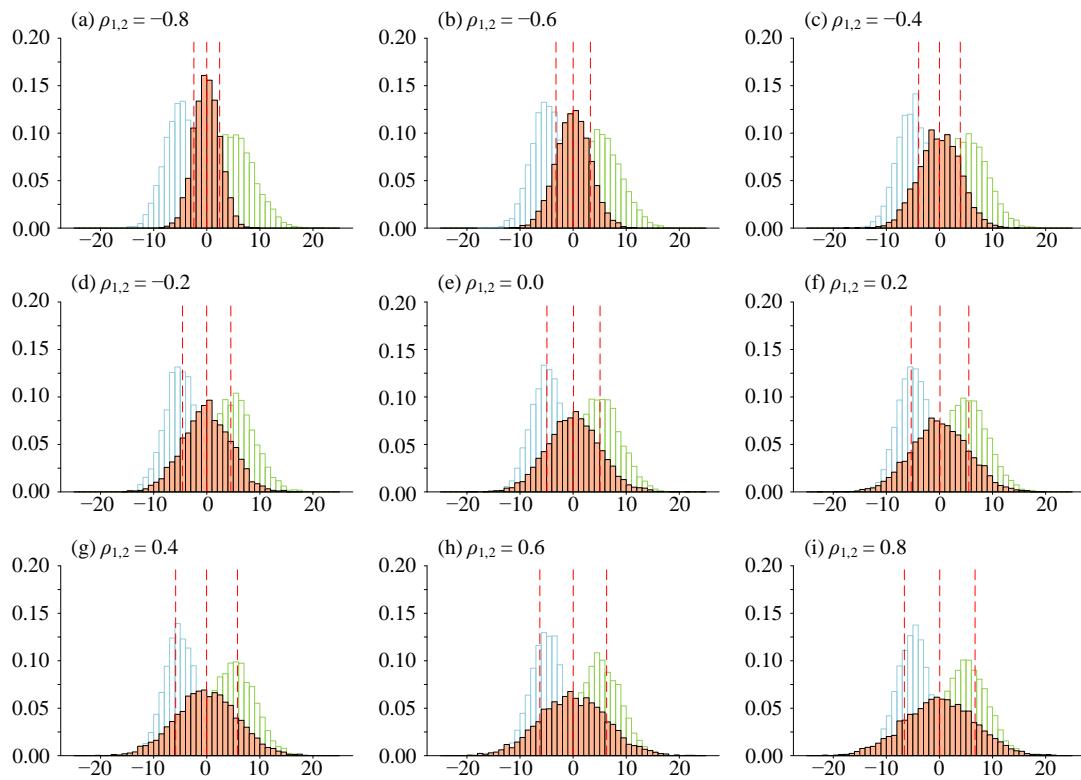
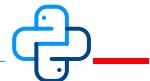
X_1 和 X_2 分布之和，即 $Y = X_1 + X_2$ ，也服从正态分布：

$$Y \sim N(\mu_Y, \sigma_Y^2) \quad (19)$$

其中，

$$\begin{aligned} \mu_Y &= \mu_1 + \mu_2 \\ \sigma_Y^2 &= \sigma_1^2 + \sigma_2^2 + 2\rho_{1,2}\sigma_1\sigma_2 \end{aligned} \quad (20)$$

图 22 所示为相关性系数 $\rho_{1,2}$ 影响 $X_1 + X_2$ 随机数分布。请大家利用几何视角分析图中不同子图结果。

图 22. 相关系数如何影响 $X_1 + X_2$ 随机数分布

Bk5_Ch15_07.py 绘制图 22。

15.10 产生满足特定相关性的随机数

Cholesky 分解



如图 23 所示，我们在《矩阵力量》第 14 章中学过如何完成“单位圆（缩放） \rightarrow 正椭圆（剪切） \rightarrow 旋转椭圆”几何变换。

经过本书上一板块的学习，大家已经清楚单位圆代表 $N(\mathbf{0}, \mathbf{I})$ ，而旋转椭圆代表 $N(\mathbf{0}, \Sigma)$ 。再经过平移，我们就可以到 $N(\mu, \Sigma)$ 。

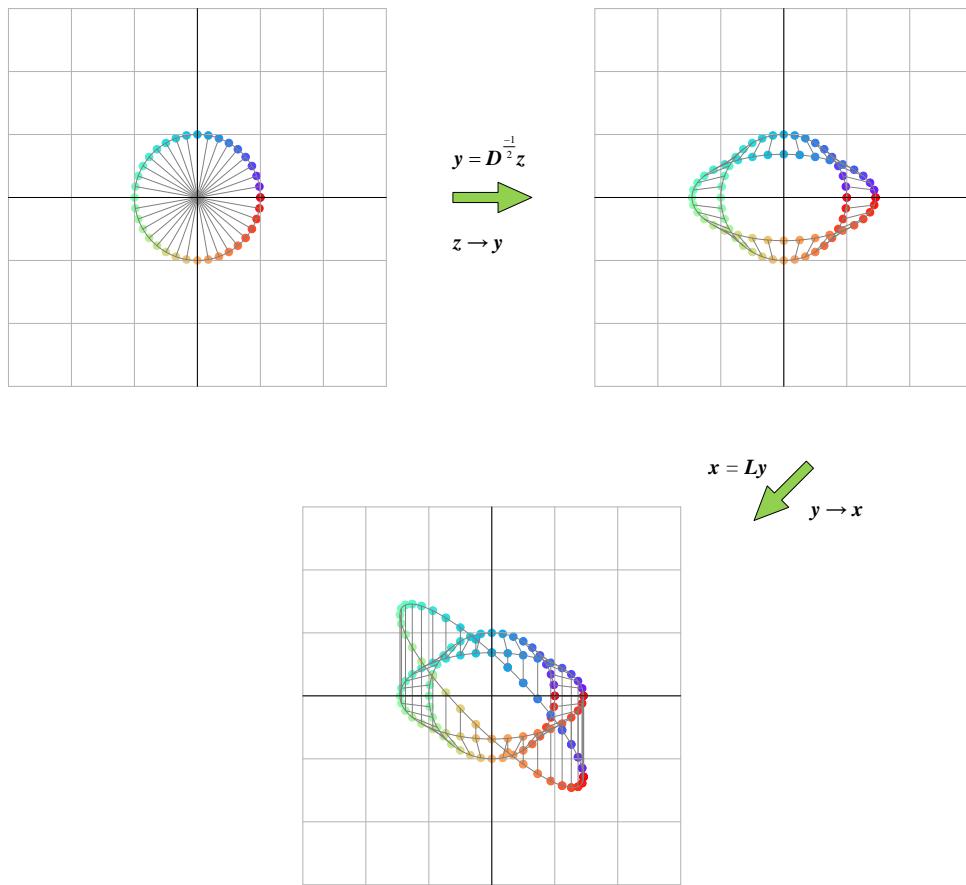


图 23. 单位圆 (缩放) → 正椭圆 (剪切) → 旋转椭圆，来自《矩阵力量》第 14 章

图 23 用到的数学工具是 LDL 分解。实际上，利用 Cholesky 分解我们可以通过一次矩阵乘法便完成“缩放 + 剪切”。这便是利用 Cholesky 分解结果产生满足特定相关性随机数的技术路线。

如图 24 所示，首先生成满足 $N(\mathbf{0}, \mathbf{I}_{D \times D})$ 的随机数矩阵 \mathbf{Z} 。

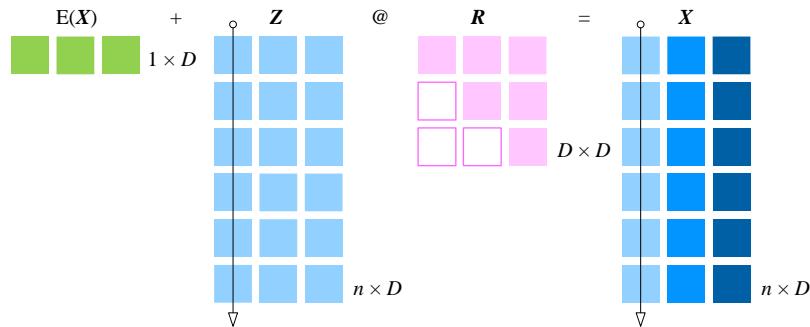


图 24. 产生满足特定相关性随机数的矩阵运算，用 Cholesky 分解结果

然后，对协方差矩阵 Σ 进行 Cholesky 分解，得到下三角矩阵 \mathbf{R}^T 和上三角矩阵 \mathbf{R} 乘积：

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

$$\Sigma = R^T R \quad (21)$$

⚠ 注意，要求 Σ 为正定；否则，不能进行 Cholesky 分解。

矩阵 R 中含有图 23 中“剪切”、“缩放”两个成分。

Z 服从 $N(\theta, I_{D \times D})$ ，经过如下运算得到的多元随机数 X 服从 $N(E(X), \Sigma_{D \times D})$ ：

$$\begin{array}{rcl} X & = & Z \\ N(E(X), \Sigma) & \xrightarrow{N(\theta, I)} & \text{Scale + shear} \\ & & \text{Translate} \end{array} \quad (22)$$

其中

$$X = [x_1 \ x_2 \ \cdots \ x_D], \quad Z = [z_1 \ z_2 \ \cdots \ z_D], \quad E(X) = [\mu_1 \ \mu_2 \ \cdots \ \mu_D] \quad (23)$$

分别计算 Z 和 X 的协方差矩阵。 Z 的协方差矩阵：

$$\Sigma_Z = \frac{Z^T Z}{n-1} = \frac{1}{n-1} \begin{bmatrix} z_1^T \\ z_2^T \\ \vdots \\ z_D^T \end{bmatrix} \begin{bmatrix} z_1 & z_2 & \cdots & z_D \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} z_1^T z_1 & z_1^T z_2 & \cdots & z_1^T z_D \\ z_2^T z_1 & z_2^T z_2 & \cdots & z_2^T z_D \\ \vdots & \vdots & \ddots & \vdots \\ z_D^T z_1 & z_D^T z_2 & \cdots & z_D^T z_D \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{D \times D} \quad (24)$$

对 X 求协方差：

$$\begin{aligned} \Sigma_X &= \frac{(X - E(X))^T (X - E(X))}{n-1} \\ &= \frac{(ZR)^T ZR}{n-1} = R^T \frac{Z^T Z}{n-1} R = R^T R = \Sigma \end{aligned} \quad (25)$$

二维随机数

下面，我们先看 $D = 2$ 这个特殊情况。

二维随机变量 χ 满足如下二维高斯分布：

$$\chi = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \underbrace{\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}}_{\Sigma}\right) \quad (26)$$

而 Z_1 和 Z_2 服从标准正态分布，且不相关。也就是说 (Z_1, Z_2) 服从 $N(\theta, I_{2 \times 2})$ ：

$$\varsigma = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\Sigma}\right) \quad (27)$$

对 (26) 协方差矩阵 Cholesky 分解：

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \underbrace{\begin{bmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{bmatrix}}_{\mathbf{L}} \underbrace{\begin{bmatrix} \sigma_1 & \rho\sigma_2 \\ 0 & \sigma_2\sqrt{1-\rho^2} \end{bmatrix}}_{\mathbf{R}} \quad (28)$$

也就是说， χ 的 ς 关系为：

$$\chi = \mathbf{L}\varsigma + \boldsymbol{\mu} \quad (29)$$

请大家思考为什么 (22) 采用上三角矩阵 \mathbf{R} ，而上式采用下三角矩阵 \mathbf{L} 。

$D=2$ 时，展开 (29) 得到：

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2\sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad (30)$$

即

$$\begin{cases} X_1 = \sigma_1 Z_1 + \mu_1 \\ X_2 = \rho\sigma_2 Z_1 + \sigma_2\sqrt{1-\rho^2} Z_2 + \mu_2 \end{cases} \quad (31)$$

下面给出一个具体示例。

图 25 (a) 给出的二维随机数满足：

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \quad (32)$$

图 25 (b) 给出的二维随机数满足：

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}\right) \quad (33)$$

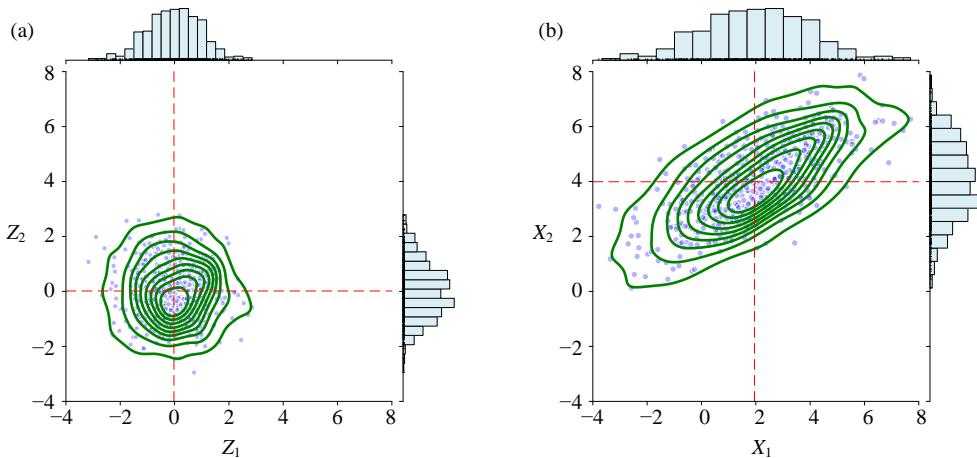
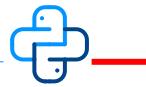


图 25. 将服从 IID 二维标准正态分布随机数转化为满足特定质心和协方差要求的随机数



Bk5_Ch15_08.py 生成图 25。代码中用到了 Cholesky 分解。

多维随机数

图 26 所示为采用多元高斯分布随机数发生器生成的随机数。这组随机数的均值、协方差矩阵和鸢尾花数据相同。请大家利用本节前文介绍的技术原理，首先生成满足 $N(\mathbf{0}, \mathbf{I}_{D \times D})$ 的随机数矩阵 \mathbf{Z} ，然后再生成满足 $N(\mathbf{E}(\mathbf{X}), \Sigma_{D \times D})$ 的随机数。

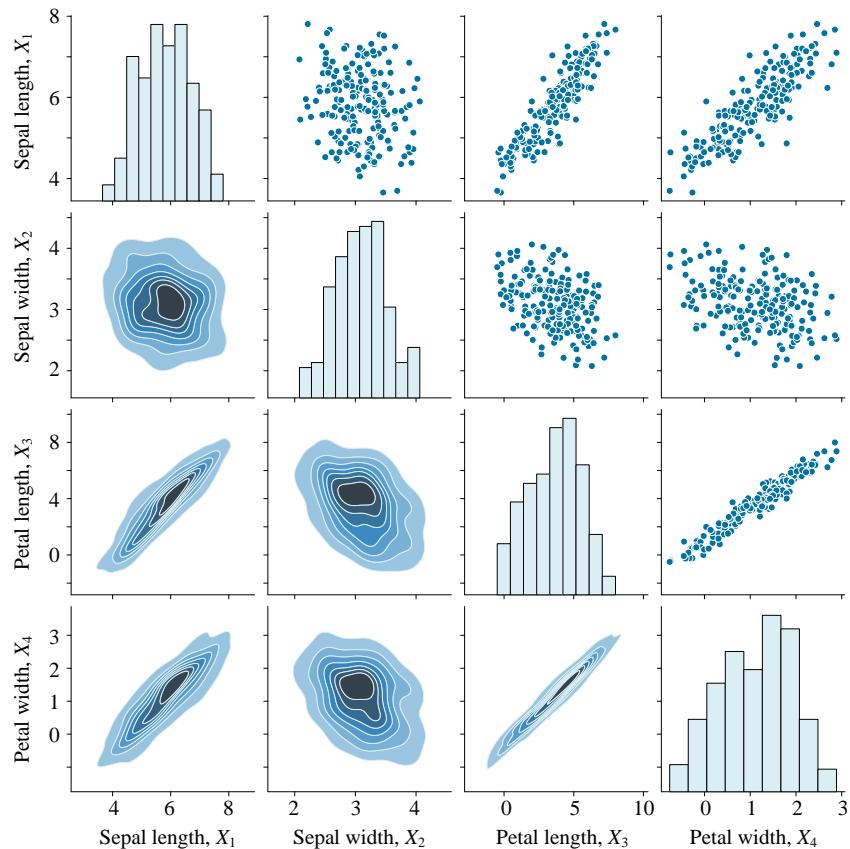
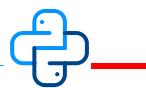


图 26. 四元高斯随机数



Bk5_Ch15_09.py 产生图 26 结果。

特征值分解

《矩阵力量》第 14 章还介绍过图 27 这幅图。图中，单位圆首先经过缩放得到正椭圆，然后正椭圆经过旋转得到旋转椭圆。这实际上是另外一条获得特定相关性随机数的技术路径。

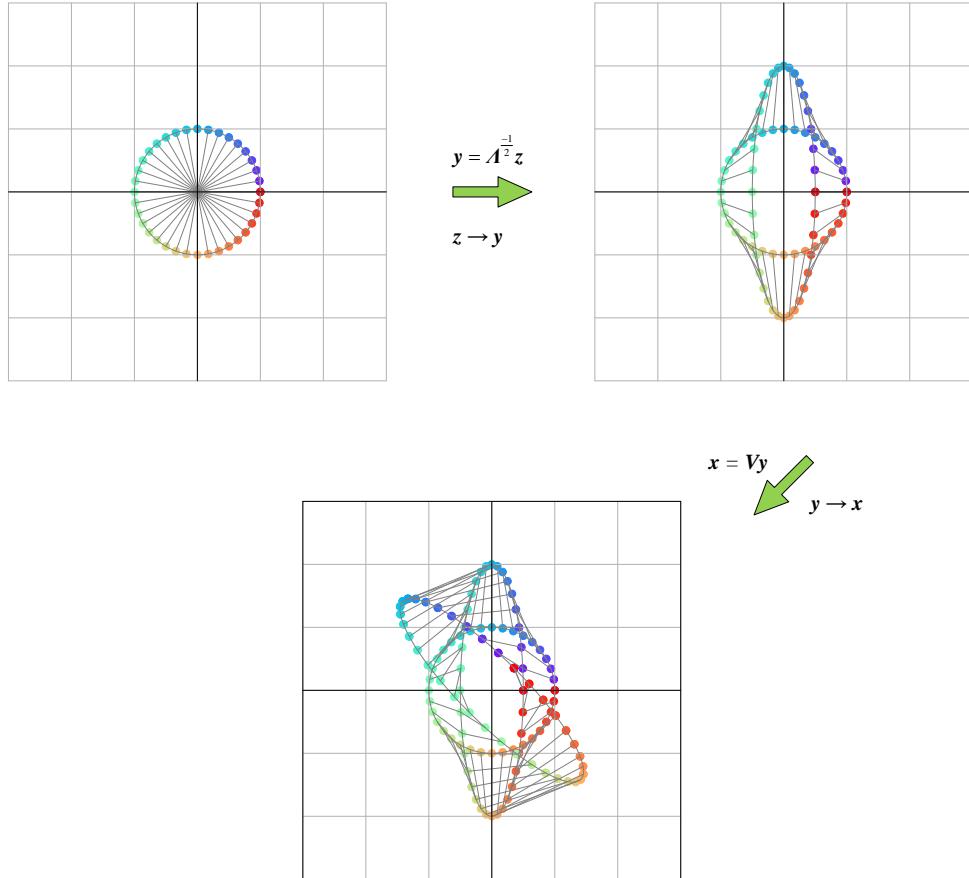


图 27. 单位圆 (缩放) → 正椭圆 (旋转) → 旋转椭圆

对协方差矩阵 Σ 进行特征值分解，然后写成“平方式”：

$$\Sigma = V \Lambda V^T = \left(\Lambda^{1/2} V^T \right)^T \Lambda^{1/2} V^T \quad (34)$$

如图 28 所示，随机数矩阵 Z 满足 $N(\theta, I_{D \times D})$ ，先经过 $\Lambda^{1/2}$ 缩放，再经过 V^T 旋转，最后通过 $E(X)$ 平移获数据矩阵 X ：

$$X = Z \Lambda^{1/2} V^T + E(X) \quad (35)$$

$N(E(X), \Sigma)$ $N(\theta, I)$ Scale Rotate Translate

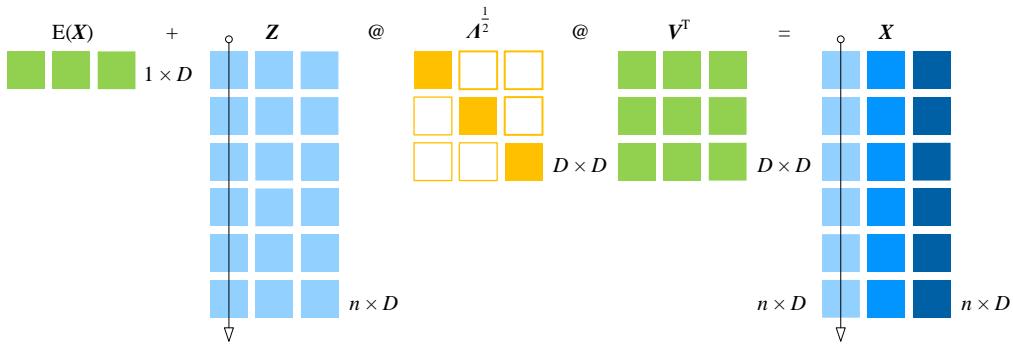


图 28. 产生满足特定相关性随机数的矩阵运算，用特征值分解结果

对 X 求协方差：

$$\begin{aligned}\boldsymbol{\Sigma}_X &= \frac{(\mathbf{X} - \mathbf{E}(\mathbf{X}))^\top (\mathbf{X} - \mathbf{E}(\mathbf{X}))}{n-1} \\ &= \frac{\left(\mathbf{Z} \mathbf{A}^{\frac{1}{2}} \mathbf{V}^T \right)^\top \mathbf{Z} \mathbf{A}^{\frac{1}{2}} \mathbf{V}^T}{n-1} = \left(\mathbf{A}^{\frac{1}{2}} \mathbf{V}^T \right)^\top \mathbf{Z}^\top \mathbf{Z} \mathbf{A}^{\frac{1}{2}} \mathbf{V}^T = \mathbf{V} \mathbf{A}^{\frac{1}{2}} \mathbf{A}^{\frac{1}{2}} \mathbf{V}^T = \boldsymbol{\Sigma}\end{aligned}\quad (36)$$

请大家利用这条技术路径生成图 25 和图 26。

一组特殊的平行四边形

对比 (22) 和 (35)，大家可能已经发现 \mathbf{R} 相当于 $\mathbf{A}^{\frac{1}{2}} \mathbf{V}^T$ 。而 \mathbf{R} 和 $\mathbf{A}^{\frac{1}{2}} \mathbf{V}^T$ 相当于协方差矩阵 $\boldsymbol{\Sigma}$ 的“平方根”。这说明协方差矩阵 $\boldsymbol{\Sigma}$ 的“平方根”不唯一。《矩阵力量》反复强调过这一点。

这意味着，凡是能够写成如下形式的矩阵 \mathbf{B} 都是协方差矩阵 $\boldsymbol{\Sigma}$ 的“平方根”：

$$\boldsymbol{\Sigma} = \mathbf{B}^T \mathbf{B} \quad (37)$$

比如，

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbf{R}^T \mathbf{R} \\ \boldsymbol{\Sigma} &= \left(\mathbf{A}^{\frac{1}{2}} \mathbf{V}^T \right)^T \mathbf{A}^{\frac{1}{2}} \mathbf{V}^T\end{aligned}\quad (38)$$

而上两式代表完全不同的几何变换。如图 29 (a) 所示，我们能够明显地看到 Cholesky 分解中的剪切操作。图 29 (b) 则明显地看出来旋转。

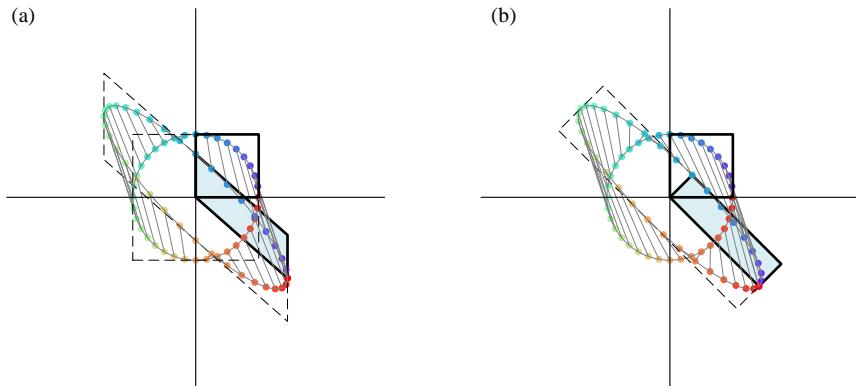


图 29. 对比 Cholesky 分解和特征值分解

更一般地，用数据矩阵 \mathbf{Z} 代表正圆，即 $N(\mathbf{0}, \mathbf{I})$ 。 \mathbf{Z} 先经过 \mathbf{U} 旋转，然后再用 \mathbf{R} 完成“缩放 + 剪切”，最后用 $E(\mathbf{X})$ 平移，得到数据矩阵 \mathbf{X} ，这个过程对应的矩阵运算为：

$$\mathbf{X} = \underset{N(\mathbf{0}, \Sigma)}{\mathbf{Z}} \underset{N(\mathbf{0}, \mathbf{I}) \text{ Rotate}}{\mathbf{U}} \underset{\text{Scale + shear}}{\mathbf{R}} + \underset{\text{Translate}}{E(\mathbf{X})} \quad (39)$$

注意， \mathbf{U} 提供旋转操作，因此 \mathbf{U} 是正交矩阵，满足 $\mathbf{U}^T \mathbf{U} = \mathbf{U} \mathbf{U}^T = \mathbf{I}$ 。

计算 \mathbf{X} 协方差矩阵，结果还是 Σ ：

$$\begin{aligned} \Sigma_{\mathbf{X}} &= \frac{(\mathbf{X} - E(\mathbf{X}))^T (\mathbf{X} - E(\mathbf{X}))}{n-1} \\ &= \frac{(\mathbf{ZUR})^T \mathbf{ZUR}}{n-1} = (\mathbf{UR})^T \frac{\Sigma_{\mathbf{Z}}}{n-1} \mathbf{UR} = \mathbf{R}^T \mathbf{U}^T \mathbf{U} \mathbf{R} = \Sigma \end{aligned} \quad (40)$$

也就是说，给定不同的旋转矩阵 \mathbf{U} ，我们就可以获得不同的 Σ 平方根 \mathbf{UR} 。也就相当于，这些完全不同的 \mathbf{UR} 都可以获得满足特定相关性条件随机数。

图 30 左上角第一幅子图实际上就是图 29 (b) 特征分解对应的几何变换。

图 30 所示为一系列不同旋转矩阵 \mathbf{U} ，在这些 \mathbf{U} 的作用下，我们最终都获得了相同的椭圆。但是仔细观察，会发现“彩灯”的运动轨迹完全不同。

旋转矩阵 \mathbf{U} 作用于单位圆，不改变单位圆的解析式。但是， \mathbf{U} 却改变了“彩灯”的位置。这实际上也回答了《矩阵力量》第 14 章有关“彩灯”位置的问题。

图 30 中一系列平行四边形都和旋转椭圆相切。相比旋转椭圆，这些平行四边形更能体现 \mathbf{UR} 的几何变换。

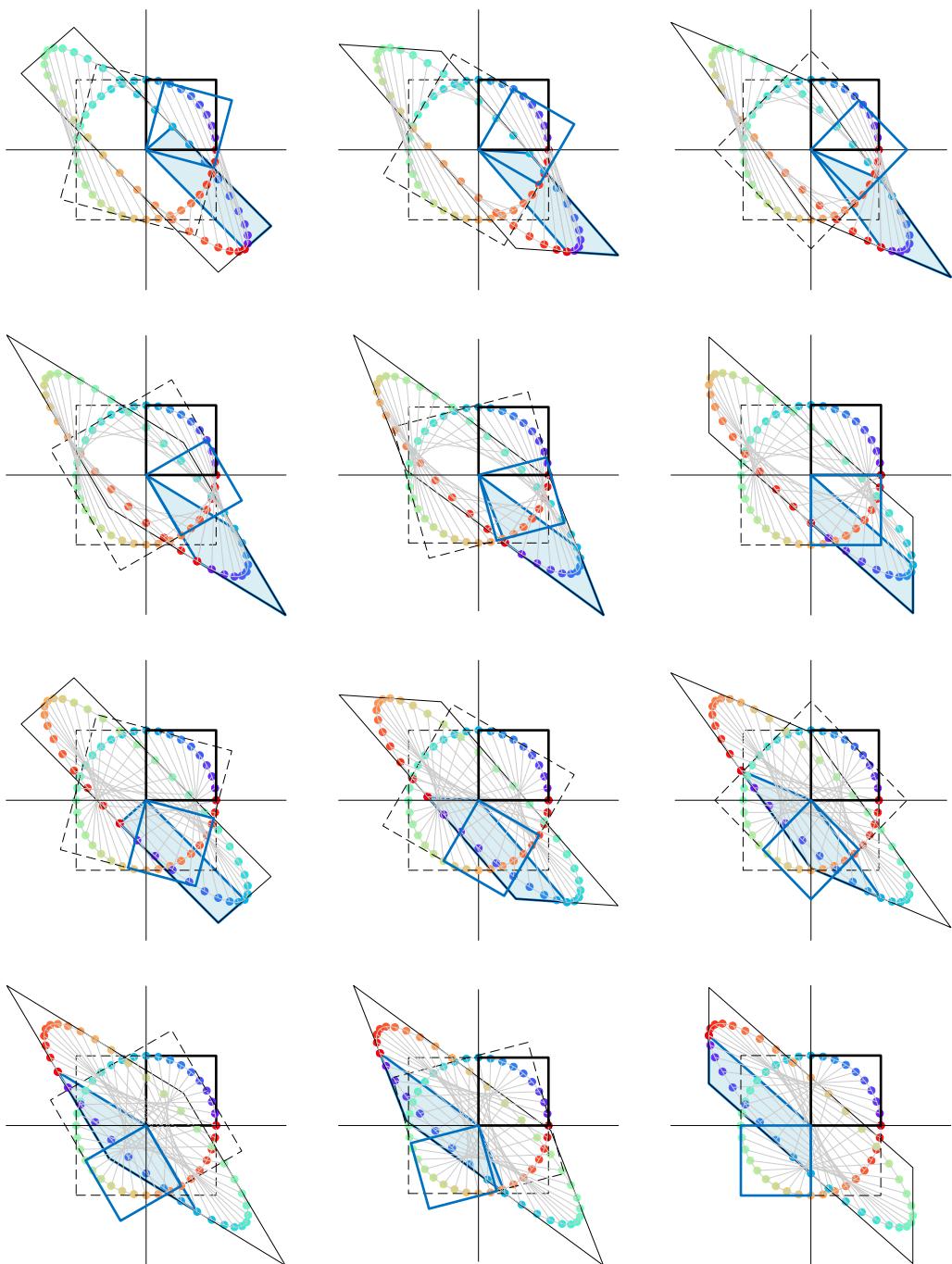


图 30. 不同的旋转矩阵



我们用 Streamlit 制作了一个应用，可视化图 30，大家可以输入不同旋转角度并绘制图 30 子图。请大家参考 Streamlit_Bk5_Ch15_10.py。



蒙特卡罗模拟的基本思想是利用随机抽样的方法来生成一组服从特定概率分布的随机数，然后用这些随机数代替原始问题中的未知量，计算问题的输出结果。通过对大量随机数进行抽样和统计，可以获得问题的近似解，从而分析问题的性质和特点。

蒙特卡罗模拟广泛应用于金融、物理、工程、生物、环境、社会科学等领域，例如金融风险评估、物理系统建模、生物统计、环境影响评价、社会网络分析等。它是一种高度灵活和通用的计算方法，可以适用于各种不同的问题和应用场景。本书后续会用马尔科夫链蒙特卡罗模拟 MCMC 完成贝叶斯推断。《数据有道》将会继续这一话题。

16

Frequentist Inference

频率派统计推断

参数固定，但不可知，将概率解释为反复抽样的极限频率



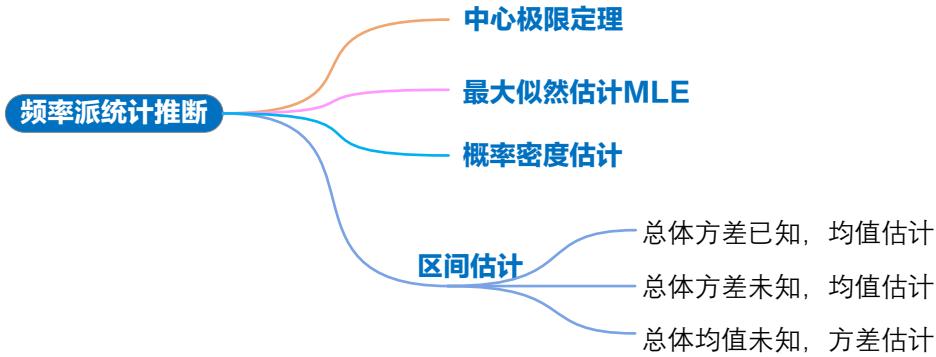
审视数学，你会发现，它不仅是颠扑不破的真理，而且是至高无上的美丽——那种冷峻而朴素的美，不需要唤起人们任何的怜惜，没有绘画和音乐的浮华装饰，纯粹，只有伟大艺术才能展现出来的严格完美。

Mathematics, rightly viewed, possesses not only truth, but supreme beauty — a beauty cold and austere, like that of sculpture, without appeal to any part of our weaker nature, without the gorgeous trappings of painting or music, yet sublimely pure, and capable of a stern perfection such as only the greatest art can show.

——伯特兰·罗素 (Bertrand Russell) | 英国哲学家、数学家 | 1872 ~ 1970



- ◀ `scipy.stats.binom_test()` 计算二项分布的 p 值
- ◀ `scipy.stats.norm.interval()` 产生区间估计结果
- ◀ `seaborn.heatmap()` 产生热图
- ◀ `seaborn.lineplot()` 绘制线型图
- ◀ `scipy.stats.ttest_ind()` 两个独立样本平均值的 t-检验



16.1 统计推断：两大学派

统计有两大分支：统计描述、统计推断。

本书第2章专门介绍了如何用图形和汇总统计量描述样本数据。而**统计推断** (statistical inference) 的数学工具来自于概率，本书“概率”、“高斯”、“随机”这三个板块给我们提供了足够的数学工具。因此，这个板块和下一板块正式进入统计推断这个话题。

本书前文提过，统计推断通过样本推断总体，在数据科学、机器学习应用颇为广泛。统计推断有两大学派——**频率学派推断** (Frequentist inference) 和**贝叶斯学派推断** (Bayesian inference)。

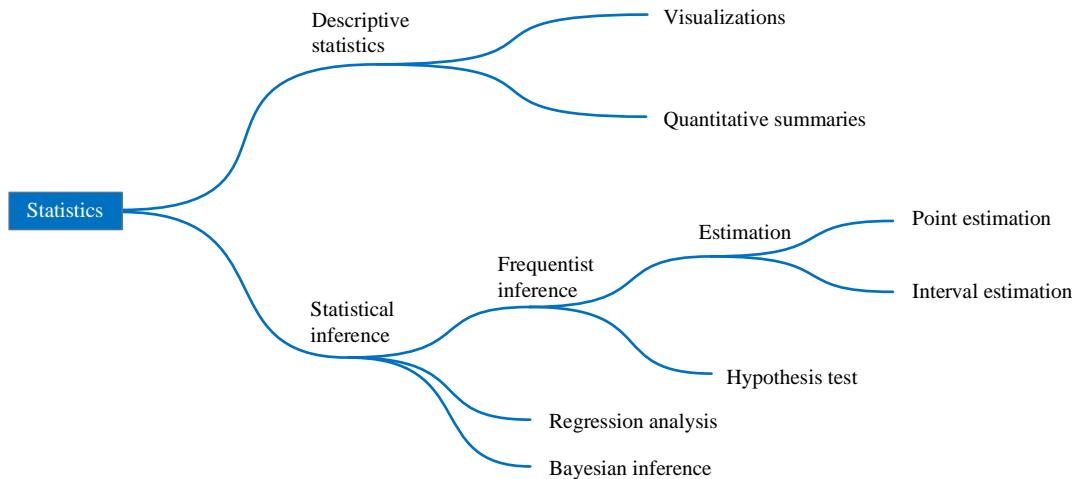


图 1. 本书统计学版图

频率学派

频率学派认为真实参数确定，但一般不可知。真实参数就好比上帝视角能够看到一切随机现象表象下的本质。

而我们观察到的样本数据都是在这个参数下产生的。真实参数对于我们不可知，频率派强调通过样本数据计算得到的频数、概率、概率密度等而得出有关总体的推断结论。

频率学派认为事件的概率是大量重复独立试验中频率的极限值。事件的可重复性、减小抽样误差对于频率派试验很重要。

频率学派方法的结论主要有两类：1) 显著性检验的“真或假”结论；2) 置信区间是否覆盖真实参数的结论。为了得出这些结论，我们需要掌握**区间估计** (interval estimation)、**最大似然估计** (maximum likelihood estimation, MLE)、**假设检验** (hypothesis test) 等数学工具。

这一章仅仅蜻蜓点水地介绍几个常用的频率学派工具，需要大家必须掌握的是最大似然估计 MLE。

⚠ 注意，本书不会介绍假设检验。《数据有道》中讲解线性回归时会涉及到常见假设检验。

贝叶斯学派

贝叶斯学派则认为参数本身也是不确定的，参数本身也是随机变量，因此也服从某种概率分布。也就是说，所有参数都可能是产生样本数据的参数，只不过不同的参数对应的概率有大有小。

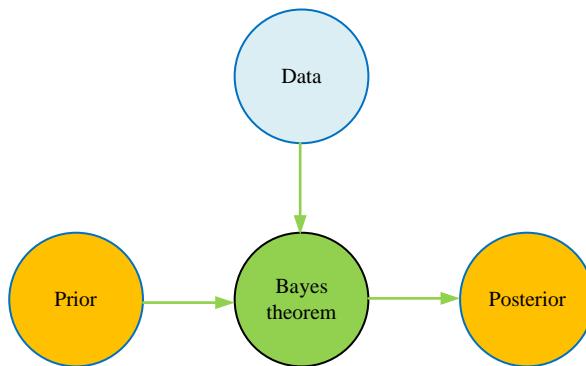


图 2. 贝叶斯推断

不同于频率派仅仅使用样本数据，贝叶斯学派结合过去的经验知识和样本数据。贝叶斯学派引入**先验分布**(prior distribution)、**后验分布**(posterior distribution)、**最大后验概率估计**(Maximum A Posteriori estimation, MAP)这样的概念来计算不同参数值的概率。

比较来看，频率学派推断只考虑证据，不考虑先验概率。频率派强调概率是可重复性事件发生的频率，而不是基于主观判断的个人信念或偏好。

此外，很多情况下，贝叶斯推断没有后验分布的解析解，因此经常利用蒙特卡罗模拟获取满足特定后验分布的随机数。本书中大家会看到 Metropolis-Hastings 抽样算法的应用。

有意思的是，当样本数据量趋近无穷时，频率学派和贝叶斯学派结果趋于一致，可谓殊途同归。

贝叶斯统计能够整合主观、客观不同来源的信息，并作出合理判断，这是频率派推断做不到的。机器学习算法中，贝叶斯统计的应用越来越广泛。

→ 本书前文提到，机器学习算法中频率学派的方法有其局限性。因此和常见的概率统计教材不同，本书“厚”贝叶斯学派，“薄”频率学派。本章和下一章将简要介绍频率学派统计推断的常用工具。而本书下一个板块将用五章内容专门介绍贝叶斯学派统计推断。

回归分析

回归分析 (regression analysis) 经常被划分到频率学派的工具箱中。作者则认为解释回归分析的视角很多，比如最小二乘优化视角、投影视角、矩阵分解、条件概率、最大似然估计 MLE、最大后验估计 MAP。因此，本书不把回归分析划在频率学派下面。

→ 本书将在第 24 章从多视角来看回归分析。另外，《数据有道》一册则有专门讲解回归分析的板块，其中大家会看到拟合优度、方差分析 ANOVA、 F 检验、 t 检验、置信区间等工具在回归分析中的应用。除了线性回归，《数据有道》还会介绍非线性回归、贝叶斯回归、基于主成分分析的回归算法。

16.2 频率学派的工具

以鸳尾花数据为例

鸳尾花数据集最初由 Edgar Anderson 于 1936 年在加拿大加斯帕半岛上采集获得。在开始本章之前，先给大家出个问题，如何设计试验估算：

- ◀ 加斯帕半岛上所有鸳尾花花萼长度均值；
- ◀ 半岛上三类鸳尾花 (setosa、versicolour、virginica) 的具体比例。

为了解决这些实际问题，统计学家想出来了两个方法来解决。

大数定理

第一个办法是尽可能多地采集样本，比如在估算加斯帕半岛上所有鸳尾花花萼长度均值时，尽量同一时间采集尽可能多的鸳尾花数据。

这里应用到的统计学原理是**大数定律** (law of large numbers)。大数定律指的是当样本数量越多时，样本的算术平均值有越大的概率接近其真实的概率分布的期望。

简单来说，大数定理告诉我们，当我们进行大量的随机实验时，随着实验次数的增加，实际观测值越来越接近真实值。这就是大数定理的“大数”之处，有点“大力出奇迹”的味道。

大数定律体现出一些随机事件的均值具有长期稳定性。本书前文提到，抛一枚硬币，硬币落地正面朝上还是反面朝上，是偶然的。但是，如果硬币质地均匀，让我们抛硬币的次数达到上千万次，就会发现硬币朝上的次数约为 50%。因此，频率学派推断特别强调同一试验的可重复性。

然而，这种办法需要尽可能多地提高样本数量，这使得试验本身变得尤为困难。

中心极限定理

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>
本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>
欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

第二种方法是，多次地独立地从总体中抽取样本，并计算每次样本的平均值，并用这些样本平均值去估算总体的期望。这种方法在统计学中被称为**中心极限定理** (central limit theorem)。

中心极限定理成立的条件包括：1) 独立性：随机变量必须是相互独立的，也就是说，一个随机变量的取值不受其他随机变量的影响。2) 相同分布：随机变量应当具有相同的概率分布，即从同一总体中独立抽取样本。3) 样本量要足够大。

中学物理课，我们用游标卡尺反复测量同一物体的厚度，然后计算平均值来估计物体的实际厚度，这一试验的思路实际上就是中心极限定理的应用。

具体来说，中心极限定理指一个总体中随机进行 n 次抽样，每次抽取 m 个样本，计算其平均数，一共能得到 n 个平均数。当 n 足够大时，这 n 个平均数的分布接近于正态分布，不管总体的分布如何。这个定理，常常也被戏谑地称为“上帝视角”，在他眼中正态分布仿佛如同宇宙终极分布一般。

游标卡尺反复测量同一物体的厚度，可能会出现一些误差。这些误差可能来自于游标卡尺的不稳定性、读数不准确、人为误差等等因素。如果我们对这些误差进行统计分析，通常可以得到一个误差分布，该分布的中心点表示这些测量的平均值，标准差表示这些测量的离散程度。

当我们进行大量的游标卡尺测量时，由于中心极限定理的作用，这些误差的分布将趋向于正态分布。因此，我们可以使用正态分布模型来描述这些误差，从而对它们进行统计分析。这些分析包括计算平均值、标准差、置信区间等，可以帮助我们评估测量结果的准确性和稳定性，以及确定测量误差的来源。

点估计

点估计 (point estimation)，顾名思义，是指用样本统计量的某单一具体数值直接作为某未知总体参数的最佳估值。

举个例子，农场有几万只鸡兔。为了估计兔子的平均体重，我们从农场动物中随机抽取 100 只兔子作为样本，计算它们的平均体重为 5 kg。如果我们选择用 5 kg 代表整个农场所有兔子的体重，这种方法就是点估计。

本章主要介绍**最大似然估计** (Maximum Likelihood Estimation, MLE)。最大似然估计 MLE 在机器学习中应用广泛，MLE 和贝叶斯学派的最大后验概率估计 MAP 地位并列。

此外，点估计也用在贝叶斯推断中。贝叶斯推断中最常用的点估计是后验分布的期望值，称为后验期望。

区间估计

在用多次抽样估计总体分布的期望时，抽样的次数总是有限的，也有可能存在极端的样本值，这都会对估算产生影响。统计学家就想到一个更有效的办法，在进行估算时将注意力集中到样本平均值可能的一个范围或区间内，并给出真实的期望值位于这个区间的概率。这个区间就被称为**置信区间** (confidence interval, CI)。

举个例子，每次抽样的次数不变，做 100 次抽样，分别计算得到 100 个对应的样本平均值，并且认定在“上帝视角”中这 100 个样本平均值服从正态分布。那么，在这个正态分布的中心区域的 95 个样本均值，就构成了一个区间。这个区间就是对应的 95% 置信区间。它告诉我们，有 95% 的可能性总体真正的期望值在这个置信区间范围内。

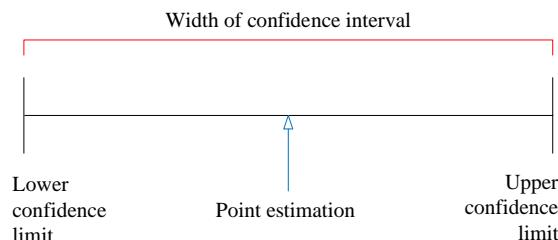


图 3. 对比点估计和区间估计

16.3 中心极限定理：渐近于正态分布

随机变量 X_1, X_2, \dots, X_n 独立同分布 IID，即相互独立且服从同一分布。 $X_k (k = 1, 2, \dots, n)$ 的期望和方差为：

$$\text{E}(X_k) = \mu, \quad \text{var}(X_k) = \sigma^2 \quad (1)$$

这 n 个随机变量的平均值 \bar{X} 近似服从如下正态分布：

$$\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (2)$$

注意，以上结论和 X_k 服从任何分布无关。

标准误 (standard error, SE) 的定义为：

$$SE = \frac{\sigma}{\sqrt{n}} \quad (3)$$

本节举两个例子来讲解中心极限定理。

离散

第一个例子是离散随机变量。

如图4所示为抛一枚色子结果 X 和对应的理论概率值。 X 服从离散均匀分布。如果每次抛 n 枚色子，这 n 个色子的结果对应 $X_1, X_2 \dots X_n$ 。然后求 n 个随机变量的平均值 \bar{X} 。根据(2)， \bar{X} 服从正态分布 $N\left(\mu, \frac{\sigma^2}{n}\right)$ 。

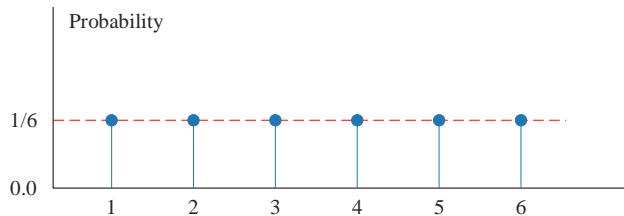
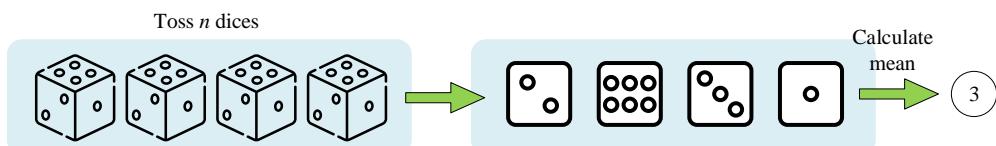


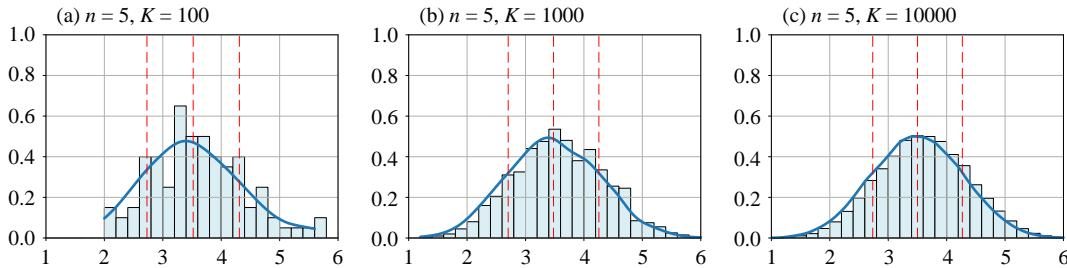
图 4. 抛一枚色子结果和对应的理论概率值

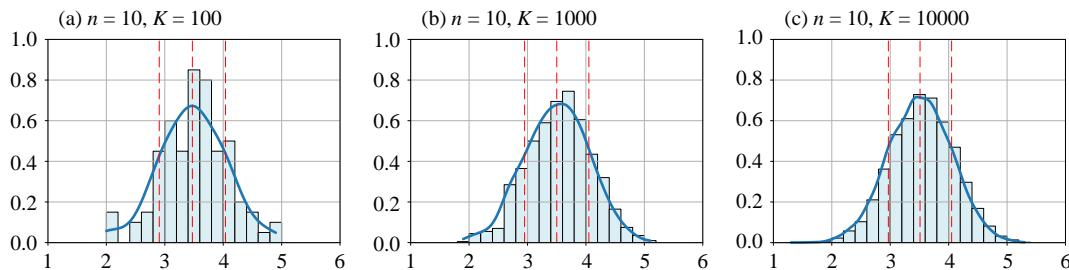
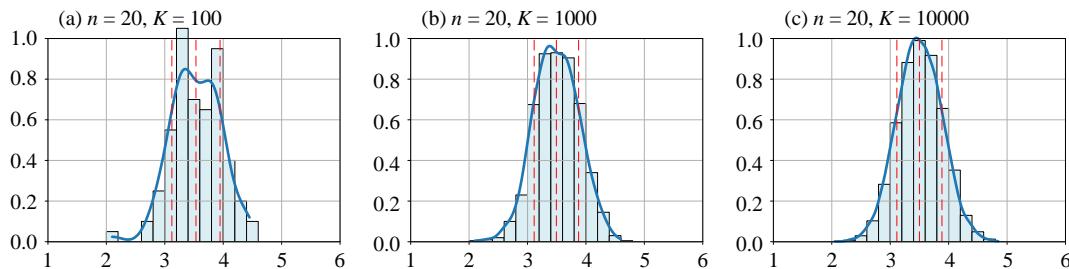
如图5所示，每次抛 n 枚色子，一共抛 K 次。下面，我们分别改变 n 和 K 进行蒙特卡罗模拟。

图 5. 每次抛 n 枚色子，一共抛 K 次

如图6所示，当 $n = 5$ 时，也就是每次抛 5 枚色子，随着 K 增大，我们很容易看出平均值 \bar{X} 趋于正态分布。

根据(2)，增大 n 会导致标准误 SE 会不断减小，对比图6、图7、图8，容易发现随着 n 增大，直方图逐渐变“瘦”，也就是说 SE 减小。

图 6. 每次抛 $n = 5$ 枚色子

图 7. 每次抛 $n = 10$ 枚色子图 8. 每次抛 $n = 20$ 枚色子

Bk5_Ch16_01.py 绘制图 6、图 7、图 8。

连续

第二个例子是连续随机变量。图 9 所示为随机数分布，这个分布有双峰，显然不是一个正态分布。如图 10 所示，试验中，每次抽取 $n = 10$ 个样本，随着试验次数 K 不断增大，平均值 \bar{X} 逐渐趋向于正态分布。图 11 中，这个趋势更加明显。图 12 所示为标准误 SE 随着 n 增大不断减小。

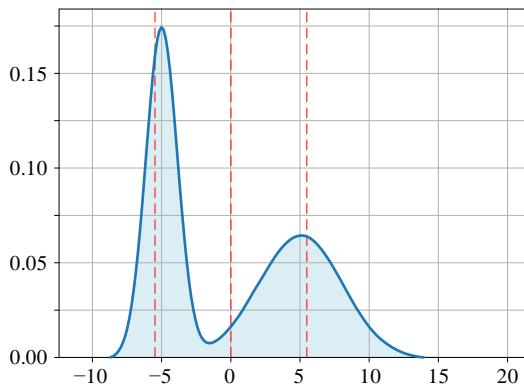


图 9. 随机数分布

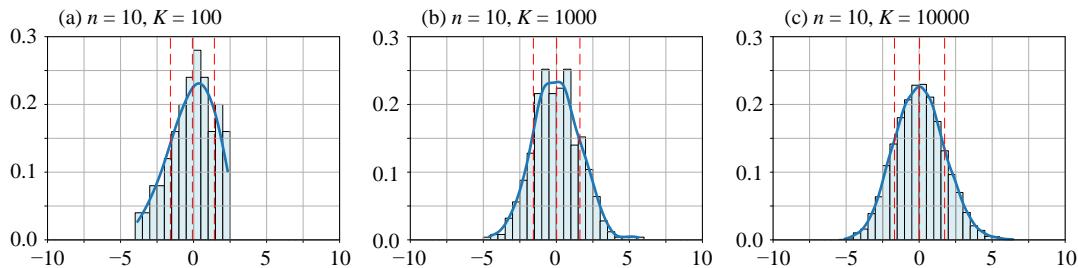


图 10. 每次抽取 10 个样本

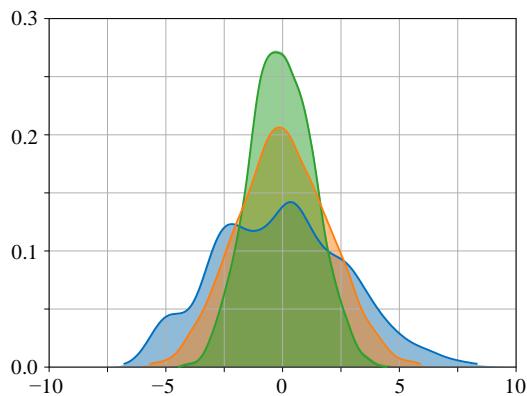


图 11. 随着试验次数增大，均值分布逐渐趋向正态

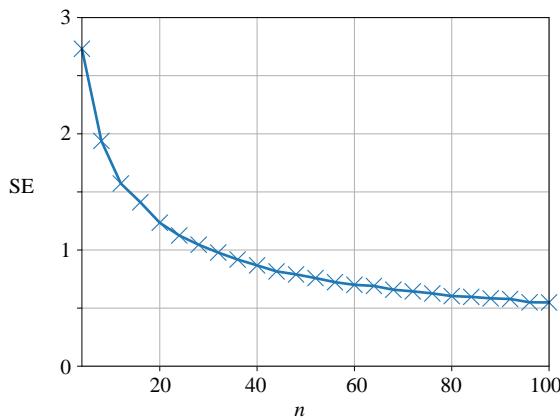


图 12. 标准误随 n 变化



Bk5_Ch16_02.py 绘制图 9、图 10、图 11、图 12。

16.4 最大似然：鸡兔比例

白话说，最大似然估计 MLE 就是找到让似然函数取得最大值的参数。

鸡兔同笼

我们先看一个简单的例子。

试想，一个农场散养大量“走地”鸡、兔。假设农场中兔子比例真实值为 θ ，但是农夫自己并不清楚。为了搞清楚农场鸡兔比例，农夫决定随机抓 n 只动物。 $X_1, X_2 \dots X_n$ 为每次抓取动物的结果。 X_i 的样本空间为 $\{0, 1\}$ ，其中 0 代表鸡，1 代表兔。

⚠ 注意，抓取动物过程，我们忽略这对农场整体动物总体比例的影响。

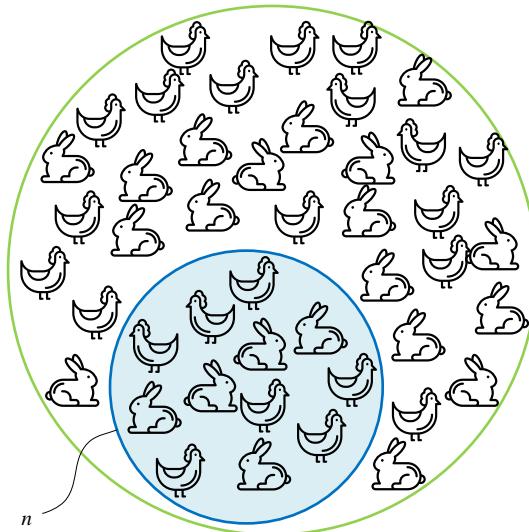


图 13. 农场有数不清的散养鸡兔

未知参数 θ

$X_1, X_2 \dots X_n$ 为 IID 的伯努利分布 $Bernoulli(\theta)$ ， X_i 的概率分布为：

$$f_{X_i}(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1-x_i} \quad (4)$$

似然函数、对数似然函数一般用 θ (theta) 作为未知量。

⚠ 注意，上式本应该是概率质量函数，但是为了方便我们还是用 $f()$ 。

⚠ 再次强调，本书前文提到过，为了避免混淆，本书用“|”引出条件概率中的条件，用分号“;”引出概率分布的参数。

似然函数

在统计学中，**似然函数** (likelihood function) 通常是通过观测数据的联合分布来定义。由于假设每个观测值都是独立同分布，所以上述联合概率可以被分解为每个观测值的边缘概率的乘积，即似然函数 $L(\theta)$ 为：

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n f_{X_i}(x_i; \theta) \\ &= \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \end{aligned} \quad (5)$$

简单来说，似然函数通常被表示为概率密度函数或概率质量函数的连乘积形式，这个连乘积表示观测数据的联合概率密度或概率质量函数。

令：

$$s = \sum_{i=1}^n x_i \quad (6)$$

s 代表 n 次抽取中兔子的总数。

这样 (5) 可以写成：

$$L(\theta) = \theta^s (1-\theta)^{n-s} \quad (7)$$

假设一次抓 20 只动物，其中 8 只兔子，则似然函数 $L(\theta)$ 为：

$$L(\theta) = \theta^8 (1-\theta)^{12} \quad (8)$$

图 14 (a) 所示为上述似然函数图像。显然，这个似然函数和横轴围成图形的面积不是 1。



本书第 20 章将介绍方法“归一化”似然函数。

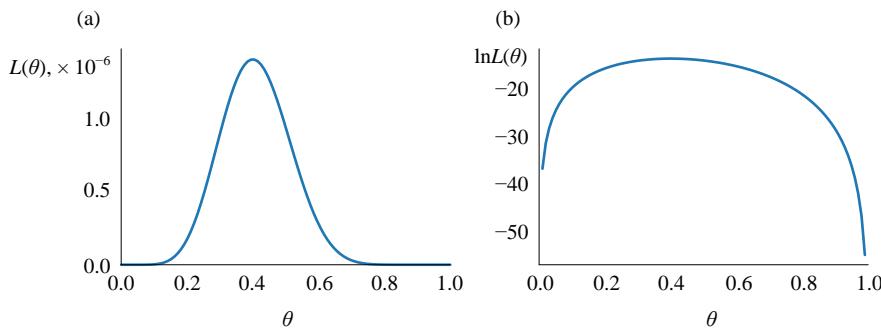


图 14. 似然函数、对数似然函数

MLE 优化问题为：

$$\arg \max_{\theta} \prod_{i=1}^n f_{X_i}(x_i; \theta) \quad (9)$$

对数似然函数

对数似然函数 (log-likelihood function) 就是对似然函数取对数，它可以将似然函数的连乘形式转换为加和形式：

$$\ln L(\theta) = s \ln \theta + (n-s) \ln(1-\theta) \quad (10)$$

当 $n = 20$, $s = 8$ 时, (10) 为：

$$\ln L(\theta) = 8 \times \ln \theta + 12 \times \ln(1-\theta) \quad (11)$$

图 14 (b) 所示为上述对数似然函数图像。



《数学要素》第 12 章提过，对数运算可以将连乘 Π 变成连加 Σ 。

在概率计算中，概率值累计乘积会经常出现数值非常小的正数情况。由于计算机的精度是有限的，无法识别这一类数据。而取对数之后，更易于计算机的识别，从而避免**浮点数下溢** (floating point underflow)。浮点数下溢，也叫**算术下溢** (arithmetic underflow)，指的是计算机浮点数计算的结果小于可以表示的最小数。

由于对数函数是单调递增的，因此最大化对数似然函数的值等价于最大化原始似然函数的值。此外，对数似然函数在计算导数时也更加方便，因为它将连乘变为加和形式，从而可以更容易地进行求导。因此，对数似然函数常常被用于最大似然估计和贝叶斯推断等统计学方法中。

优化问题

有了对数似然函数，(18) 中的 MLE 优化问题可以写成：

$$\arg \max_{\theta} \sum_{i=1}^n \ln f_{X_i}(x_i; \theta) \quad (12)$$

(10) 中 $\ln L(\theta)$ 对 θ 求偏导为 0, 构造等式:

$$\frac{d \ln L}{d \theta} = \frac{s}{\theta} - \frac{n-s}{1-\theta} = 0 \quad (13)$$

求解上式得到:

$$\hat{\theta}_{MLE} = \frac{s}{n} \quad (14)$$



我们将在本书第 21 章用贝叶斯派统计推断重新求解这个问题。

16.5 最大似然：以估算均值、方差为例

设 $X \sim N(\mu, \sigma^2)$, μ 和 σ^2 为未知参数。

X_1, X_2, \dots, X_n 来自 X 的 n 个样本, 显然 X_1, X_2, \dots, X_n 独立同分布。 x_1, x_2, \dots, x_n 是 X_1, X_2, \dots, X_n 的观察值。下面介绍利用最大似然方法求解 μ 和 σ^2 的估计量。

X_i 的概率密度函数为:

$$f_{X_i}(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma^2} \exp\left(\frac{-1}{2 \sigma^2} \left(x_i - \mu\right)^2\right) \quad (15)$$

未知参数 θ

令 $\theta_1 = \mu, \theta_2 = \sigma^2$, X_i 的概率密度函数则写成:

$$f_{X_i}(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(\frac{-1}{2\theta_2} (x_i - \theta_1)^2\right) \quad (16)$$

似然函数

似然函数 $L(\theta_1, \theta_2)$ 为 $f_{X_i}(x_i; \theta_1, \theta_2)$ 的连乘:

$$\begin{aligned}
 L(\theta_1, \theta_2) &= f_{X_1}(x_1; \theta_1, \theta_2) \cdot f_{X_2}(x_2; \theta_1, \theta_2) \cdots f_{X_n}(x_n; \theta_1, \theta_2) \\
 &= \prod_{i=1}^n f_X(x_i; \theta_1, \theta_2) \\
 &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(\frac{-1}{2\theta_2}(x_i - \theta_1)^2\right)
 \end{aligned} \tag{17}$$

对数似然函数

对 (17) 取对数得到 $\ln L(\theta_1, \theta_2)$:

$$\ln L(\theta_1, \theta_2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\theta_2) - \frac{1}{2\theta_2} \left(\sum_{i=1}^n (x_i - \theta_1)^2 \right) \tag{18}$$

优化问题

为了最大化 (18) 中 $\ln L(\theta_1, \theta_2)$, 对 θ_1, θ_2 求偏导为 0, 构造等式:

$$\begin{aligned}
 \frac{\partial \ln L}{\partial \theta_1} &= \frac{1}{\theta_2} \left(\sum_{i=1}^n (x_i - \theta_1) \right) = 0 \\
 \frac{\partial \ln L}{\partial \theta_2} &= -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2} \left(\sum_{i=1}^n (x_i - \theta_1)^2 \right) = 0
 \end{aligned} \tag{19}$$

可以求得:

$$\begin{aligned}
 \hat{\theta}_1 &= \frac{\sum_{i=1}^n x_i}{n} = \bar{X} \\
 \hat{\theta}_2 &= \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n}
 \end{aligned} \tag{20}$$

“戴帽子”的 $\hat{\theta}_1$ 、 $\hat{\theta}_2$ 为对真实 θ_1 、 θ_2 的估计。注意，上式中 $\hat{\theta}_2$ 并不是对方差的无偏估计。

具体值

给定样本为 $\{-2.5, -5, 1, 3.5, -4, 1.5, 5.5\}$, 下面用 MLE 估算其均值和方差。

将样本代入 (18), 得到对数似然函数:

$$\ln L(\theta_1, \theta_2) = -6.432 - 3.5 \ln \theta_2 - \frac{7\theta_1^2 + 93}{2\theta_2} \tag{21}$$

$\ln L(\theta_1, \theta_2)$ 对 θ_1, θ_2 求偏导为 0, 构造等式:

$$\begin{aligned}\frac{\partial \ln L}{\partial \theta_1} &= -\frac{7\theta_1}{\theta_2} = 0 \\ \frac{\partial \ln L}{\partial \theta_2} &= \frac{7\theta_1^2 - 7\theta_2 + 93}{2\theta_2^2} = 0\end{aligned}\tag{22}$$

求解上式得到：

$$\begin{aligned}\hat{\theta}_1 &= 0 \\ \hat{\theta}_2 &= 13.2857\end{aligned}\tag{23}$$

并计算得到对数似然函数的最大值：

$$\max \{ \ln L(\theta_1, \theta_2) \} = -18.98598\tag{24}$$

图 15 所示为 $\ln L(\theta_1, \theta_2)$ 曲面， \times 对应对数似然函数最大值点位置。



本书第 24 章中，我们将用到 MLE 估算线性回归参数。

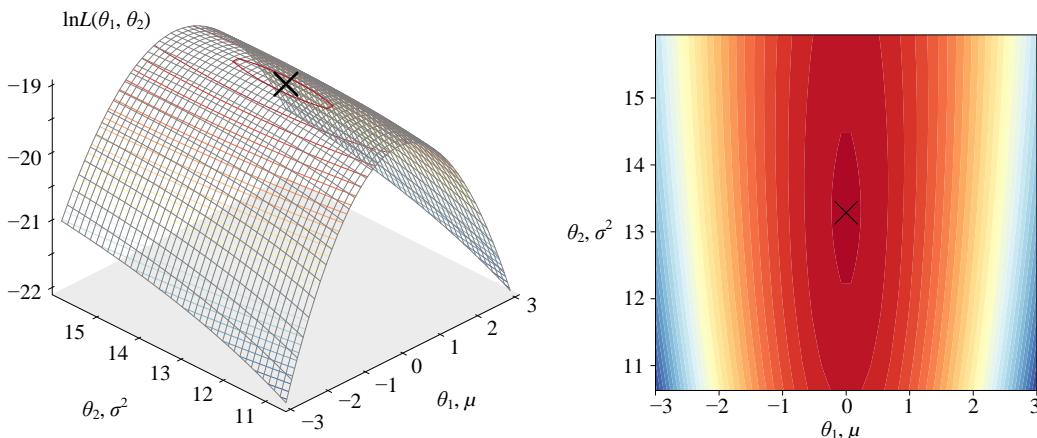
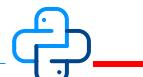


图 15. $\ln L(\theta_1, \theta_2)$ 曲面和最大值点位置



Bk5_Ch16_03.py 绘制图 15。

16.6 区间估计：总体方差已知，均值估计

不同于点估计仅给出一个数值，**区间估计** (interval estimate) 在推断总体参数时，根据统计量的抽样分布特征，估算出总体参数的一个区间范围，并且估算出总体参数落在这一区间的概率。

区间估计在点估计的基础上附加**误差限** (margin of error) 来构造**置信区间** (confidence interval)，置信区间对应的概率，被称为**置信度** (confidence level)。

本节介绍总体方差 σ^2 已知，计算给定置信水平下均值的区间估计。

双边置信区间

对于样本数据 $\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(n)}\}$ ，计算**样本平均值** (sample mean 或 empirical mean)：

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x^{(i)} \quad (25)$$

如果总体的方差已知，总体平均值 μ 的 $1 - \alpha$ 水平的**双边置信区间** (two tailed confidence interval) 可以表达为：

$$\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \quad (26)$$

其中：

\bar{X} 为**样本均值** (sample mean)。

n 为**样本数量** (sample size)。

α 为**显著性水平** (significance level)，代表的意义是在一次试验中小概率事物发生的可能性大小。 α 通常取 0.1 或 0.05。

$1 - \alpha$ 为**置信水平** (confidence level)，表示真值在置信区间内的可信程度。

$z_{1-\alpha/2}$ 叫**临界值** (critical value)，本质上就是 Z 分数。 $z_{1-\alpha/2}$ 可以通过标准正态分布的逆累计概率密度分布函数计算。

σ 为**总体的标准差** (volatility of the population)。

如图 16 所示， $1 - \alpha$ 为置信水平意味着：

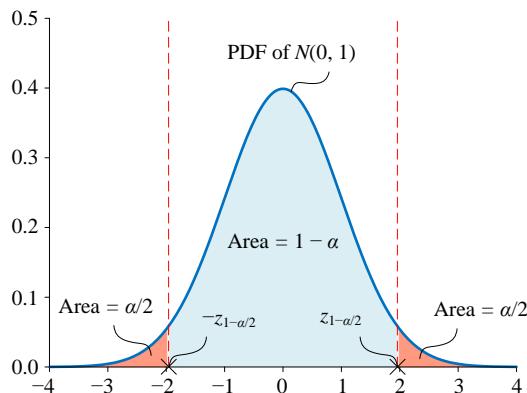
$$\Pr\left(\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (27)$$

求解 $z_{1-\alpha/2}$ 的方法为：

$$z_{1-\alpha/2} = F_{N(0,1)}^{-1}\left(1 - \frac{\alpha}{2}\right) = -F_{N(0,1)}^{-1}\left(\frac{\alpha}{2}\right) \quad (28)$$

$F_{N(0,1)}^{-1}(\cdot)$ 是标准正态分布的**逆累计分布函数** (inverse cumulative distribution function, ICDF)。

这和本书前文介绍的百分点函数 PPF 本质上一致。

图 16. 标准正态分布和 $1 - \alpha$ 置信水平

95%置信水平

总体方差已知，95% ($1 - \alpha = 1 - 5\%$) 置信水平的双边置信区间约为：

$$\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \quad (29)$$

也就是说：

$$\Pr \left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \approx 0.95 \quad (30)$$

再次强调区间估计得到的是总体参数落在某一区间的概率。图 17 (a) 所示为 100 次估算得到的 95% 置信水平的双尾置信区间。图中，黑色竖线为总体均值所在位置。

× 代表每次估算样本均值所在位置。当总体均值落在双尾置信区间时，区间为蓝色；否则，区间为红色。图 17 (a) 给出的 100 个区间中，有 88 个双尾区间包含真实的总体均值；12 个双尾区间不包含真实的总体均值。图 17 (b) 为每次抽取得样的样本数据分布山脊图。

增大每次抽样样本数量 n ，左侧置信区间不断收窄，而右侧分布范围不断变宽，两者并不矛盾。请大家思考背后原因。

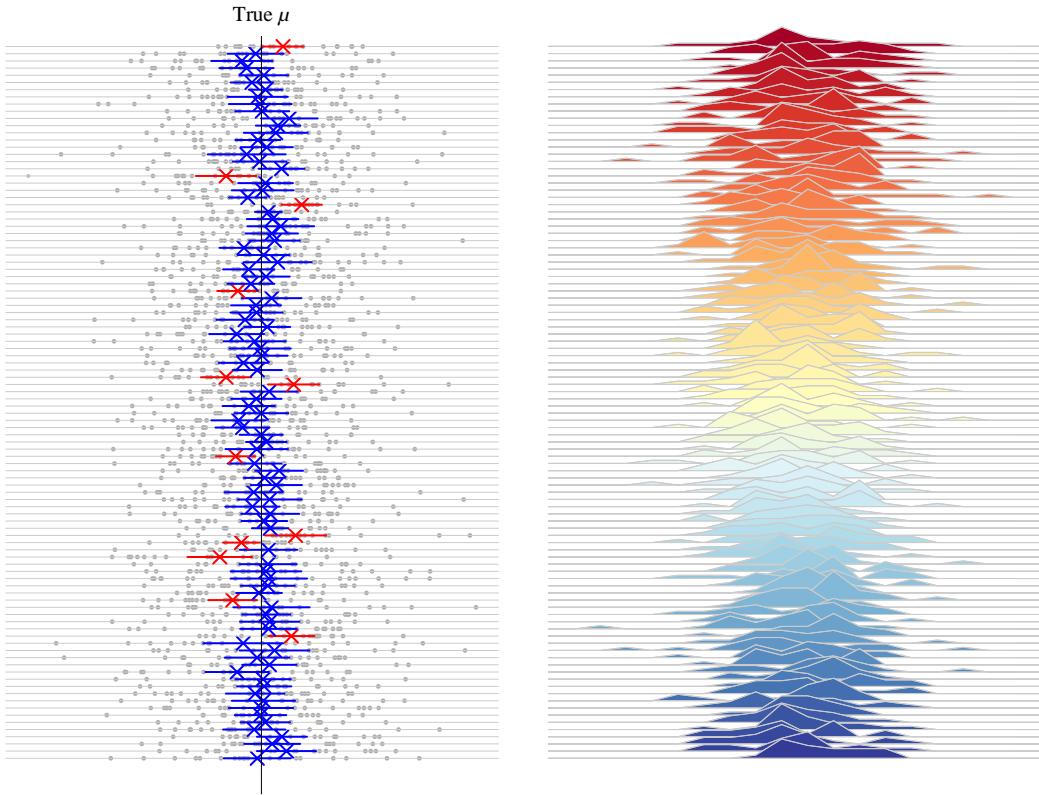


图 17. 100 次估算得到的 95% 置信水平的双尾置信区间，每次数据的分布的山脊图

单边置信区间

除了双边置信区间，统计上还常用**单边置信区间** (one-tailed confidence interval)。单边置信区间可以“左尾”，即取值范围从负无穷到平均值 \bar{X} 右侧的临界值：

$$\left(-\infty, \bar{X} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) \quad (31)$$

这意味着：

$$\Pr\left(\mu < \bar{X} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (32)$$

单边置信区间也可以是“右尾”，取值范围从 \bar{X} 左侧的临界值到正无穷：

$$\left(\bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, +\infty\right) \quad (33)$$

这意味着：

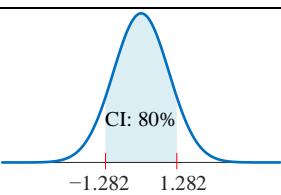
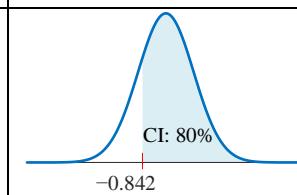
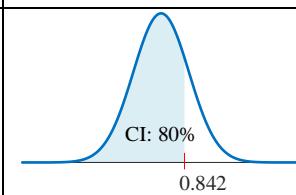
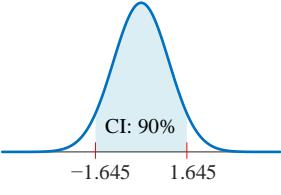
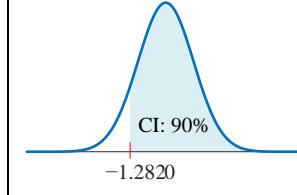
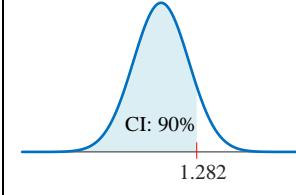
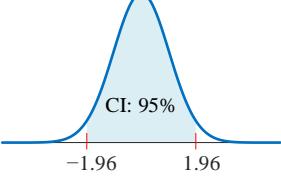
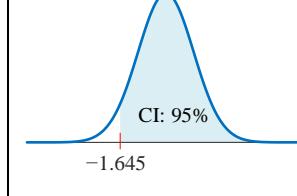
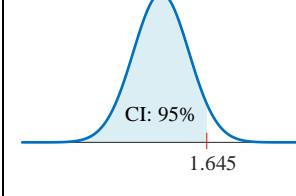
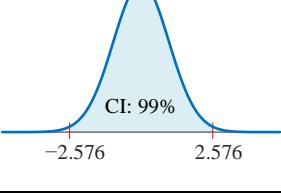
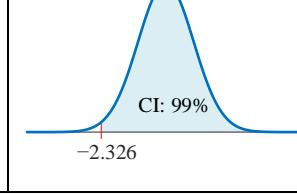
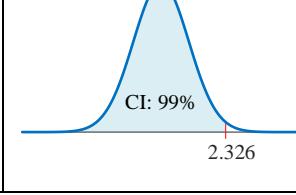
$$\Pr\left(\mu > \bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad (34)$$

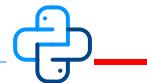
举个例子，总体方差已知，95% ($1 - \alpha = 1 - 5\%$) 水平的单侧置信区间分别为：

$$\left(-\infty, \bar{X} + 1.645 \frac{\sigma}{\sqrt{n}}\right), \quad \left(\bar{X} - 1.645 \frac{\sigma}{\sqrt{n}}, +\infty\right) \quad (35)$$

表 1 所示为不同显著性水平的双尾、左尾、右尾置信区间。

表 1. 不同显著性水平的置信区间

显著性水平 置信水平	双尾	左尾	右尾
$\alpha = 20\%$ $1 - \alpha = 80\%$			
$\alpha = 10\%$ $1 - \alpha = 90\%$			
$\alpha = 5\%$ $1 - \alpha = 95\%$			
$\alpha = 1\%$ $1 - \alpha = 99\%$			



Bk5_Ch16_04.py 绘制表 1 中图像。

16.7 区间估计：总体方差未知，均值估计

如果总体方差 σ^2 未知，就不能用上一节的估算方法。

首先，计算样本方差 s^2 ：

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{X})^2 \quad (36)$$

样本均方差 s 为：

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{X})^2} \quad (37)$$

如果总体的方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的**双边置信区间** (two tailed confidence interval) 为：

$$\left(\bar{X} - t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1) \frac{s}{\sqrt{n}} \right) \quad (38)$$

其中， n 为样本数量； $t_{1-\alpha/2}(n-1)$ 为自由度 $n-1$ ，CDF 值为 $1 - \alpha/2$ 的学生-t 的逆累计分布值。图 18 所示为自由度为 5 时， $1 - \alpha$ 置信水平双尾置信区间对应位置。

自由度较小时，学生-t 分布有明显的厚尾现象。由于厚尾现象的存在，同样的置信区间，学生 t 分布的临界值的绝对值要大于标准正态分布。但是当自由度 $df = n - 1$ 不断提高，学生-t 分布逐渐接近标准正态分布。

图 19 所示为总体方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的左尾/右尾置信区间。

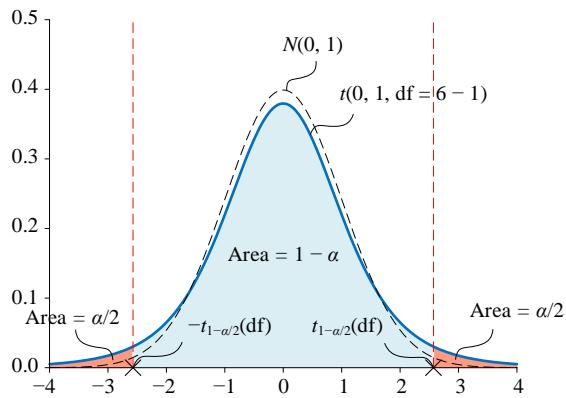
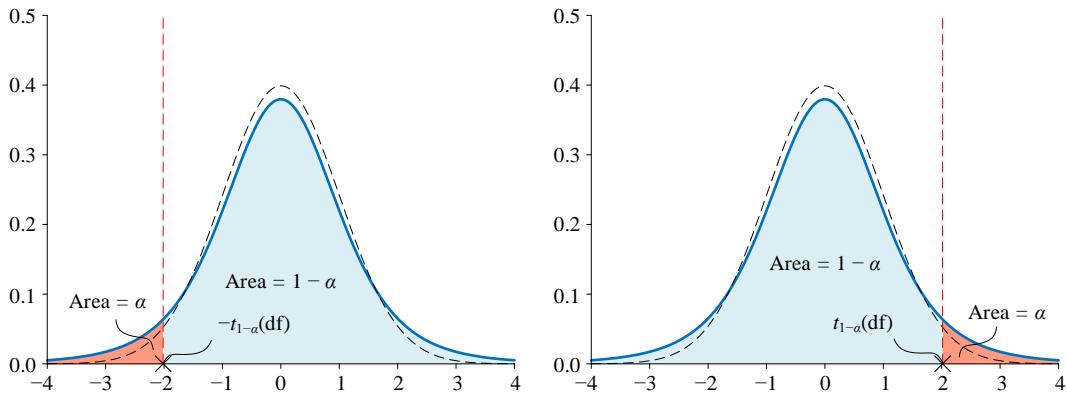
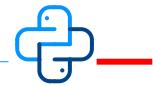


图 18. 总体方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的双尾置信区间

图 19. 总体方差未知，总体平均值 μ 的 $1 - \alpha$ 置信水平的左尾/右尾置信区间

Bk5_Ch16_05.py 绘制图 18 和图 19。

16.8 区间估计：总体均值未知，方差估计

总体均值未知情况下， σ^2 的无偏估计为 s^2 ：

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x^{(i)} - \bar{X})^2 \quad (39)$$

方差 σ^2 的 $1 - \alpha$ 水平的双边置信区间 (two tailed confidence interval) 为：

$$\left(\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)}, \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \right) \quad (40)$$

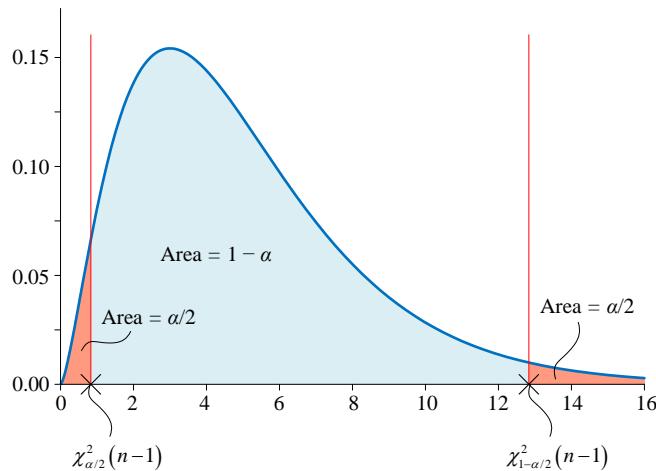
其中， n 为样本数量； $\chi_{\alpha/2}^2(n-1)$ 为自由度 $n-1$ 的卡方分布。我们还会在本书第 23 章有关马氏距离的内容用到卡方分布。

(40) 意味着：

$$\Pr\left(\frac{(n-1)s^2}{\chi_{1-\alpha/2}^2(n-1)} < \sigma^2 < \frac{(n-1)s^2}{\chi_{\alpha/2}^2(n-1)} \right) = 1 - \alpha \quad (41)$$

上式开方，得到标准差 σ 的 $1 - \alpha$ 水平的双边置信区间可以表达为：

$$\left(\frac{\sqrt{n-1}s}{\sqrt{\chi_{1-\alpha/2}^2(n-1)}}, \frac{\sqrt{n-1}s}{\sqrt{\chi_{\alpha/2}^2(n-1)}} \right) \quad (42)$$

图 20 所示为总体均值未知，方差估计的 $1 - \alpha$ 置信水平的双尾置信区间。图 20. 总体均值未知，方差估计的 $1 - \alpha$ 置信水平的双尾置信区间

Bk5_Ch16_06.py 绘制图 20。



本书首先比较了统计推断的两大学派——频率学派、贝叶斯学派。频率学派认为概率是事件发生的频率，以样本为基础进行推断；而贝叶斯学派则将概率视为主观信念的度量，以先验知识为基础进行推断。两者的不同在于对概率的定义和解释方式，但两者也可以相互补充。

然后，我们简单地了解了常用的频率学派数学工具。再次说明，《统计至简》一册轻频率学派，重贝叶斯学派。这是因为机器学习、深度学习中贝叶斯学派的思想、方法、工具戏份十足。

下一章聊一聊另外一个机器学习中常用的频率学派工具——概率密度估计。



有关如何用 Python 完成假设检验，请大家自学 Stanford 这门统计学课程相关内容。

https://web.stanford.edu/class/stats110/notes/Chapter6/Large_sample.html

这门课程网站还有大量 Python 和概率统计相结合的实例，很适合初学者参考。

17

Probability Density Estimation

概率密度估计

核密度估计就是若干概率密度函数加权叠合



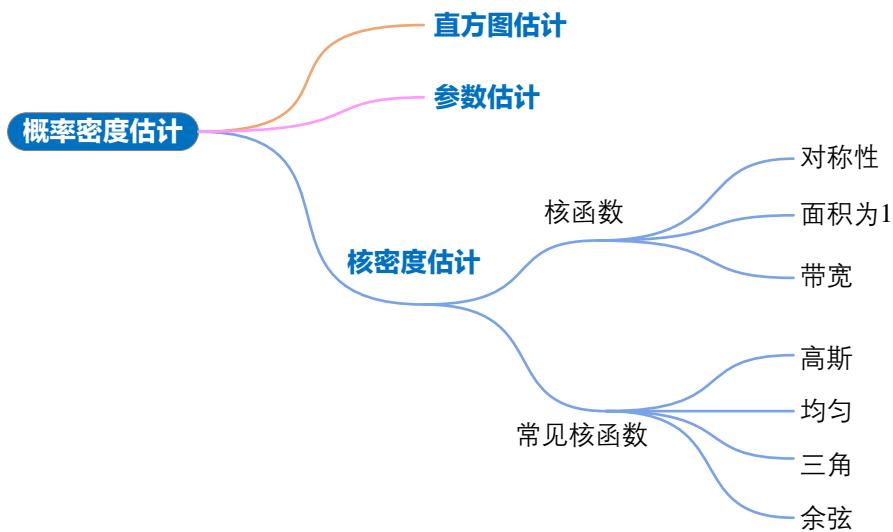
大自然是一个无限的球体，其中心无处不在，圆周无处可寻。

Nature is an infinite sphere of which the center is everywhere and the circumference nowhere.

—— 布莱兹·帕斯卡 (Blaise Pascal) | 法国哲学家、科学家 | 1623 ~ 1662



- ◀ `matplotlib.pyplot.fill_between()` 区域填充颜色
- ◀ `seaborn.kdeplot()` 绘制 KDE 概率密度估计曲线
- ◀ `sklearn.neighbors.KernelDensity()` 概率密度估计函数
- ◀ `statsmodels.api.nonparametric.KDEUnivariate()` 构造一元 KDE
- ◀ `statsmodels.nonparametric.kde.kernel_switch()` 更换核函数
- ◀ `statsmodels.nonparametric.kernel_density.KDEMultivariate()` 构造多元 KDE



17.1 概率密度估计：从直方图说起

简单来说，**概率密度估计** (probability density estimation) 就是寻找合适的随机变量概率密度函数，使其尽量贴合样本数据分布情况。

直方图

直方图实际上是最常用的一种概率密度估计方法。本书第 2 章介绍过，为了构造直方图，首先将样本数据的取值范围分为一系列左右相连等宽度的**组** (bin)，然后统计每个组内样本数据的频数。绘制直方图时，以组距为底边、以频数为高度，绘制一系列矩形图。

图 1 所示为鸢尾花四个特征上样本数据的频数直方图。合理地选择组距，让大家一眼能够通过直方图看出样本分布的大致情况。纵轴的频数，也可以替换成概率、概率密度。当纵轴为概率密度时，直方图这些矩形面积之和为 1，对应概率 1。

但是，直方图的缺点也很明显，概率密度估计结果呈现阶梯状，不“平滑”。很多数据科学、机器学习应用场合，我们需要得到连续平滑的密度估计曲线。

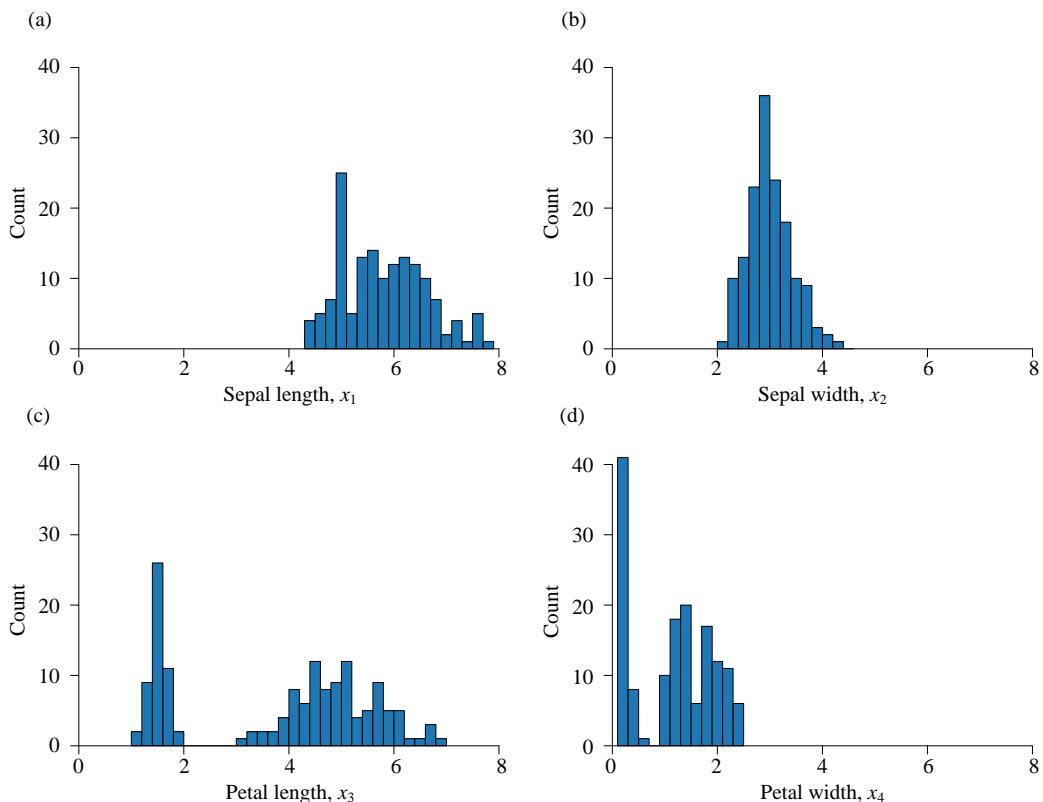


图 1. 鸢尾花四个特征的直方图，纵轴为频数

参数估计

本书前文介绍一些常见的概率分布函数，但是它们的形状远远不够描述现实世界采集的分布情况较为复杂的样本数据。

以高斯分布为例，我们可以很容易计算得到样本数据的均值 μ 和均方差 σ ，这样可以直接用正态分布来估计样本数据在某个单一特征上的分布情况：

$$\hat{f}_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad (1)$$

估计概率密度时，直接利用均值 μ 和均方差 σ 这两个参数，因此这种方法也被称作参数估计。如图 2 所示，高斯分布显然比图 1 的直方图“平滑”的多。

这种方法的缺陷是显而易见的，对比图 1 和图 2，容易发现样本分布细节被忽略，最明显的是鸢尾花花瓣长度（比较图 1 (c)、图 2 (c)）、花瓣宽度（比较图 1 (d)、图 2 (d)）这两个特征上样本数据的分布。多数情况，样本数据分布不够“正态”，仅仅使用均值 μ 和均方差 σ 描述数据不合适。

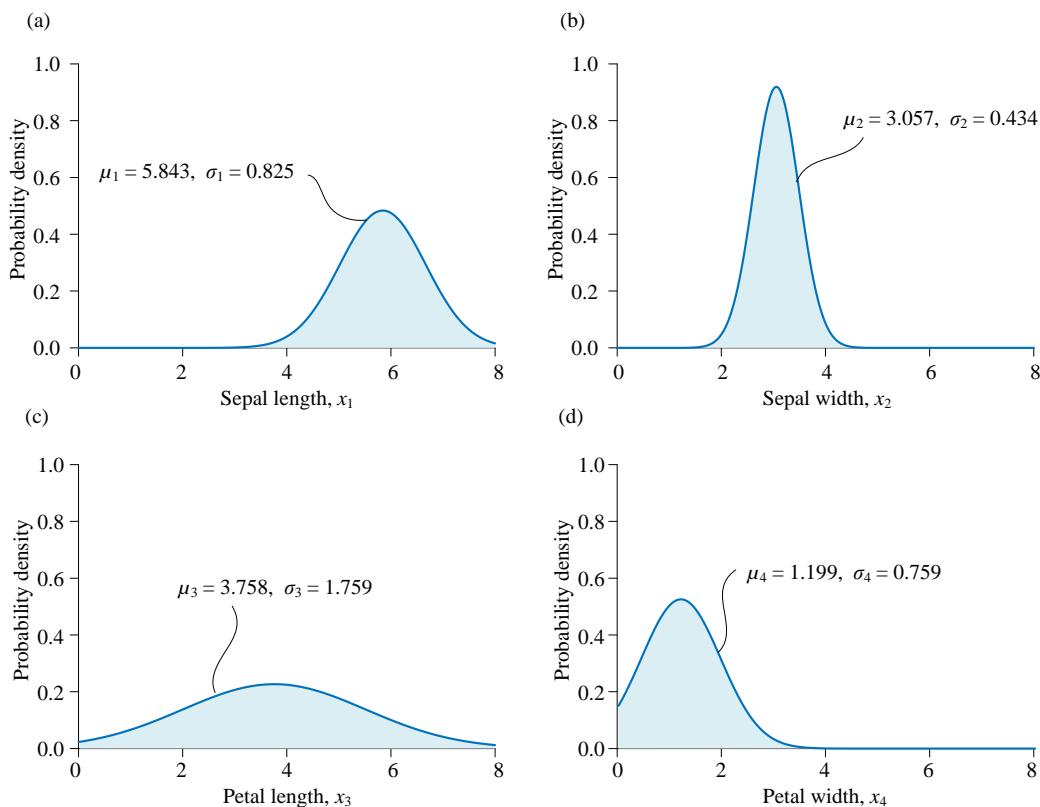


图 2. 用一元高斯分布估计鸢尾花四个特征的概率密度曲线

核密度估计

下面介绍本章的主角——**核密度估计** (Kernel Density Estimation, KDE)。本书前文很多场合已经用过核密度估计，比如第 2、5 章中都用高斯核密度估计过鸢尾花单一特征概率密度，以及联合概率密度。

核密度估计需要指定一个核函数来描述每一个数据点，最常见的核函数时高斯核函数，本章还会介绍并比较其他核函数。

图 3 所示为通过高斯核函数核密度估计得到的平滑曲线，下面我们聊一聊核密度估计原理。

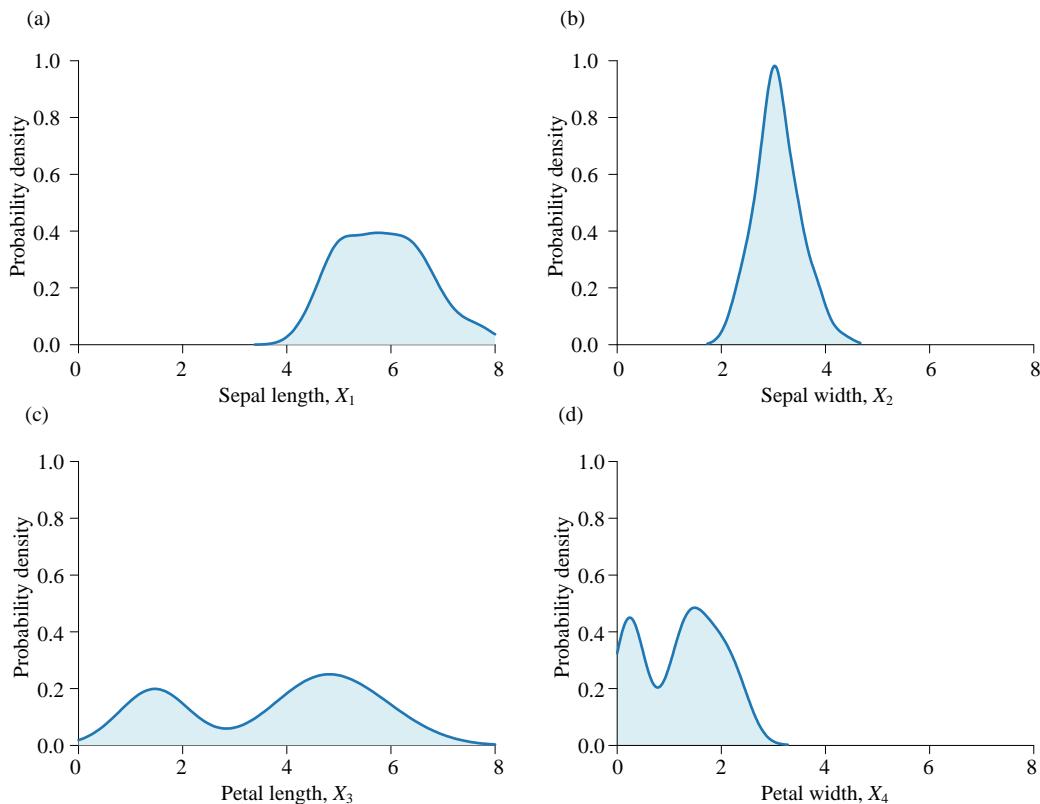


图 3. 鸢尾花四个特征的高斯 KDE 曲线



Bk5_Ch17_01.py 代码绘制图 3。代码使用 `seaborn.kdeplot()` 绘制 KDE 曲线。本章后续分别介绍几种不同的办法绘制 KDE 曲线。

17.2 核密度估计：若干核函数加权叠合

核密度估计其实是对直方图的一个自然拓展。直方图不够平滑，我们引入合适的核函数得到更加平滑的概率密度估计曲线。前文说到，核函数种类很多，本节以高斯核函数为例介绍核密度估计原理。

原理

任意一个数据点 $x^{(i)}$ ，都可以用一个函数来描述，这个函数就是核函数。如图 4 所示，一共有 7 个样本点，每一个样本点都用一个高斯核函数描述。白话说，图 4 中这 7 条曲线等权重叠加便得到核密度估计概率密度曲线。

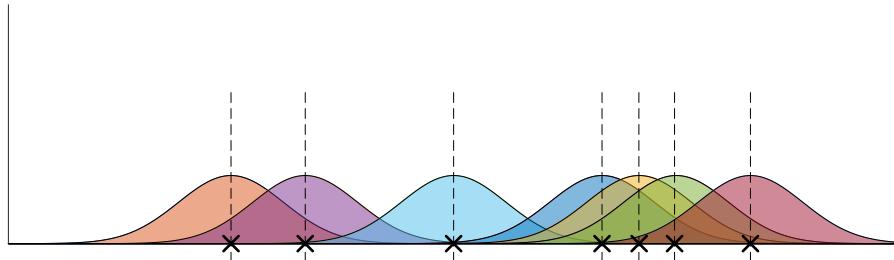


图 4. 用多个核函数描述样本数据

叠加 → 平均

而对于 n 个样本数据点 $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ ，我们可以用 n 个核函数分别代表每个数据点：

$$\underbrace{\frac{1}{h} K\left(\frac{x - x^{(i)}}{h}\right)}_{\text{Area} = 1}, \quad -\infty < x < +\infty \quad (2)$$

其中， $h (h > 0)$ 是核函数本身的缩放系数，又叫带宽。每个核函数和水平面构成图形的面积为 1。

这 n 个核函数先叠加，然后再平均，便得到概率密度估计函数：

$$\hat{f}_x(x) = \frac{1}{n} \sum_{i=1}^n K_h\left(x - x^{(i)}\right) = \underbrace{\frac{1}{n}}_{\text{Weight}} \underbrace{\frac{1}{h} \sum_{i=1}^n K\left(\frac{x - x^{(i)}}{h}\right)}_{\text{Area} = n}, \quad -\infty < x < +\infty \quad (3)$$

上式中， $1/n$ 让 n 个面积为 1 的函数面积归一化。也就是说，每个核函数贡献的面积为 $1/n$ 。

高斯核函数

下面我们以高斯核函数为例，聊聊如何理解(2)。

高斯核函数 $K(x)$ 的定义：

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x^2}{2}\right) \quad (4)$$

上述高斯核函数显然和横轴围成的面积为 1。

对称性

核函数要求具有对称性，即：

$$K(x) = K(-x) \quad (5)$$

显然，(4) 定义的高斯核函数满足对称性。

而(2) 中 $x - x^{(i)}$ 代表曲线在水平方向平移。由于核函数 $K(x)$ 关于纵轴对称，因此 $K(x - x^{(i)})$ 关于 $x = x^{(i)}$ 对称。

缩放

(2) 中的带宽 h 则代表图像在水平方向的缩放。大家是否还记得图 5？我们在讲解函数图像变换时提过，原函数 $f(x)$ 和 $cf(cx)$ 面积相同，其中 $c > 0$ 。



图 5 这两幅子图来自《数学要素》第 12 章。

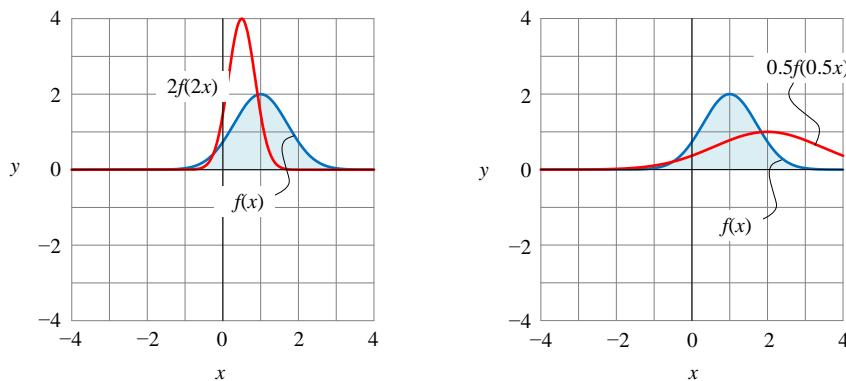


图 5. 原函数 $y = f(x)$ 水平方向、竖直方向伸缩，图片来自《数学要素》第 12 章

面积为 1

$K(x)$ 的重要性质之一是面积为 1，也就是 $K(x)$ 对 x 在 $(-\infty, +\infty)$ 积分为 1：

$$\int_{-\infty}^{+\infty} K(x) dx = 1 \quad (6)$$

(4) 中高斯核函数显然满足这一条件。

利用换元积分，很容易得到如下等式：

$$\int_{-\infty}^{+\infty} K(x) dx = \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x}{h}\right) dx = 1 \quad (7)$$

上式解释了为什么 $f(x)$ 和 $cf(cx)$ 面积相同。

可视化“叠加”

以图 4 为例，假设 7 个样本数据构成的集合为 $\{-3, -2, 0, 2, 2.5, 3, 4\}$ 。

如果 $h = 1$ ，参考 (3)，可用高斯核函数构造概率密度估计函数：

$$\hat{f}_x(x) = \frac{1}{7} \left(\frac{e^{\frac{-(x+3)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{\frac{-(x+2)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}} + \frac{e^{\frac{-(x-2)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{\frac{-(x-2.5)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{\frac{-(x-3)^2}{2}}}{\sqrt{2\pi}} + \frac{e^{\frac{-(x-4)^2}{2}}}{\sqrt{2\pi}} \right) \quad (8)$$

如图 6 所示，每个数据点给总的概率密度曲线估计贡献一条曲线。每一条曲线和横轴的面积为 $1/7$ 。叠加得到的曲面和横轴围成图形的面积为 1。

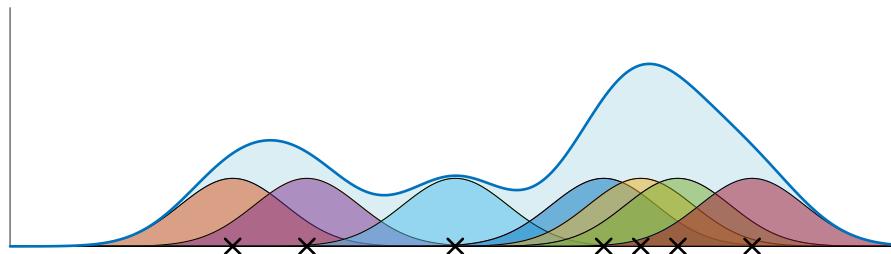


图 6. 用 7 个高斯核函数构造得到的概率密度估计曲线

以鸢尾花数据为例

图 7 所示为利用 `statsmodels.api.nonparametric.KDEUnivariate()` 对象得到的概率密度估计曲线。也可以通过它获得如图 8 所示累积概率密度估计曲线。

下一节将讲解带宽 h 如何影响概率密度估计曲线。

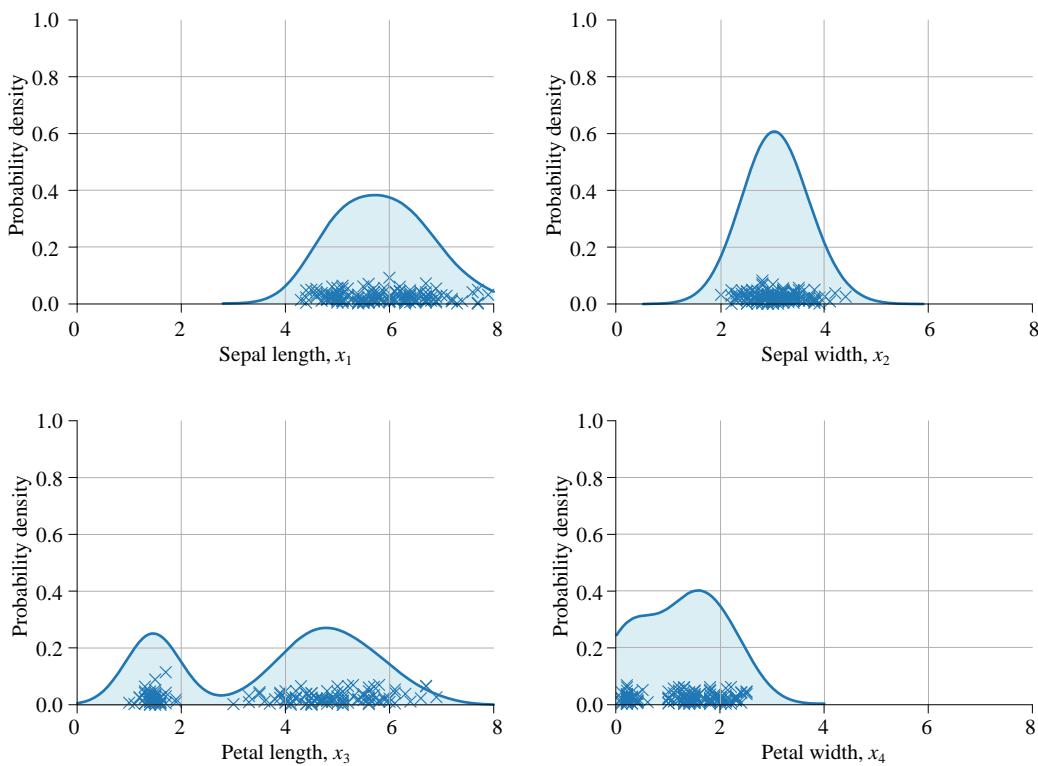


图 7. 鸢尾花四个特征数据的概率密度函数曲线

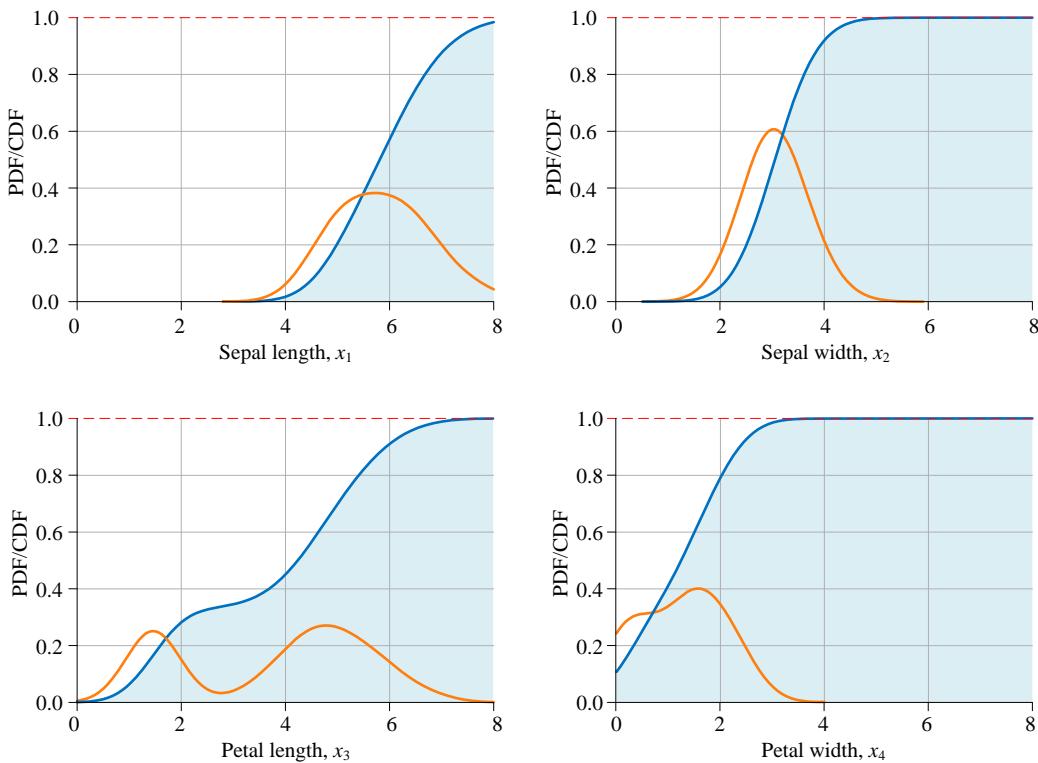


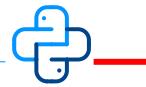
图 8. 鸢尾花四个特征数据的累积概率密度函数曲线

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



Bk5_Ch17_02.py 代码绘制图 7 和图 8。大家可以自行改变代码中带宽 h 。

17.3 带宽：决定核函数高矮胖瘦

带宽 h 选取对概率密度估计函数至关重要。 h 决定了每一个核函数的高矮胖瘦。图 9 所示为带宽 h 对高斯核函数形状影响。简单来说， h 小，核函数细高； h 大，核函数矮胖。

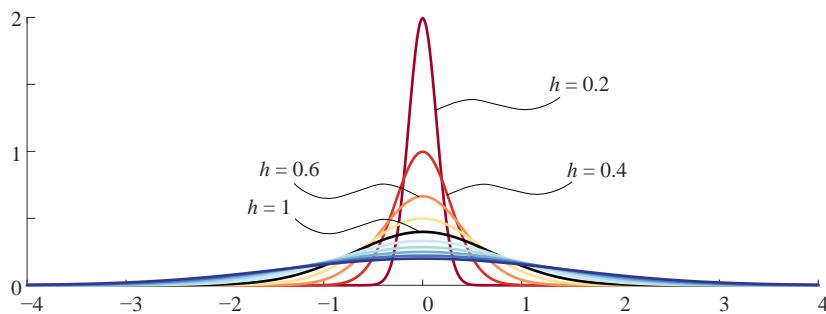


图 9. 带宽 h 对高斯核函数形状的影响

如图 10 所示，过小的 h ，会让概率密度估计曲线不够平滑；而太大的 h ，会让概率密度曲线过于平滑，大量有用信息被忽略。

⚠ 注意，不管 h 的大小，合成得到的概率密度曲线横轴包裹区域的面积始终保持为 1。

图 11 和图 12 分别展示 $h = 0.1, 1$ 时鸳尾花概率密度估计曲线。

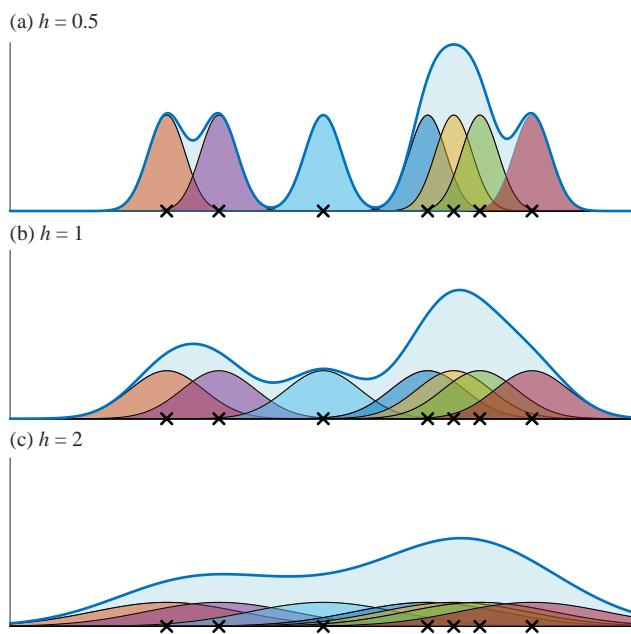


图 10. 核函数带宽对概率密度估计曲线影响

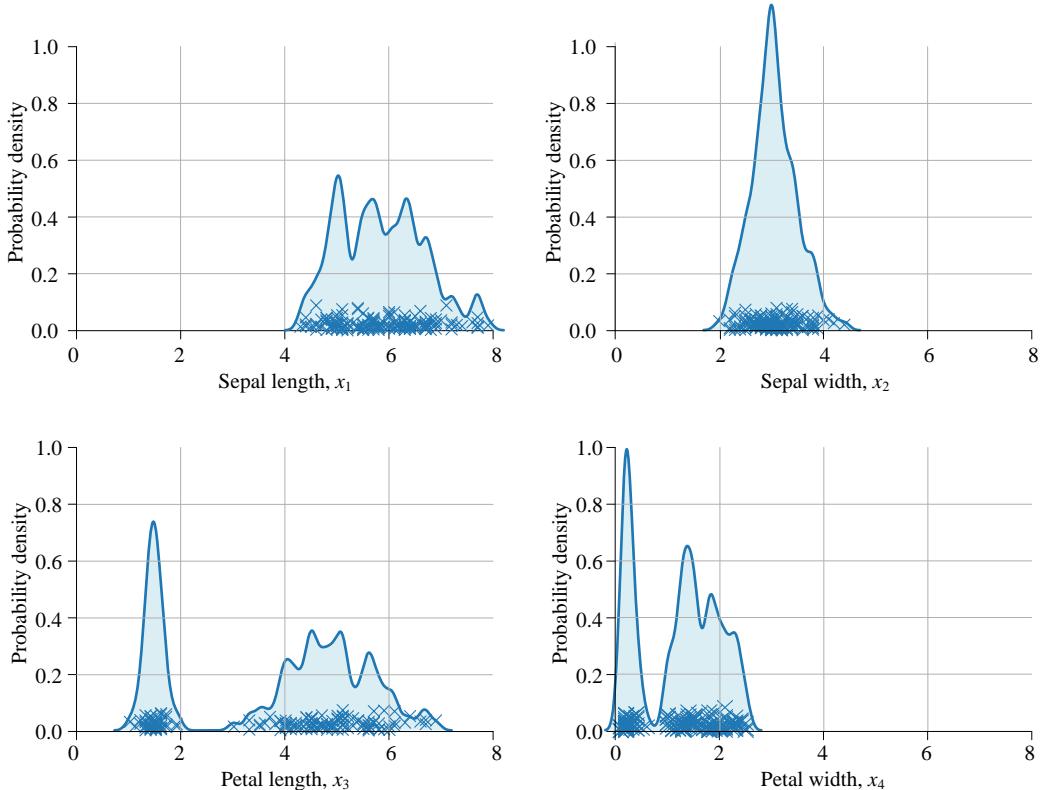
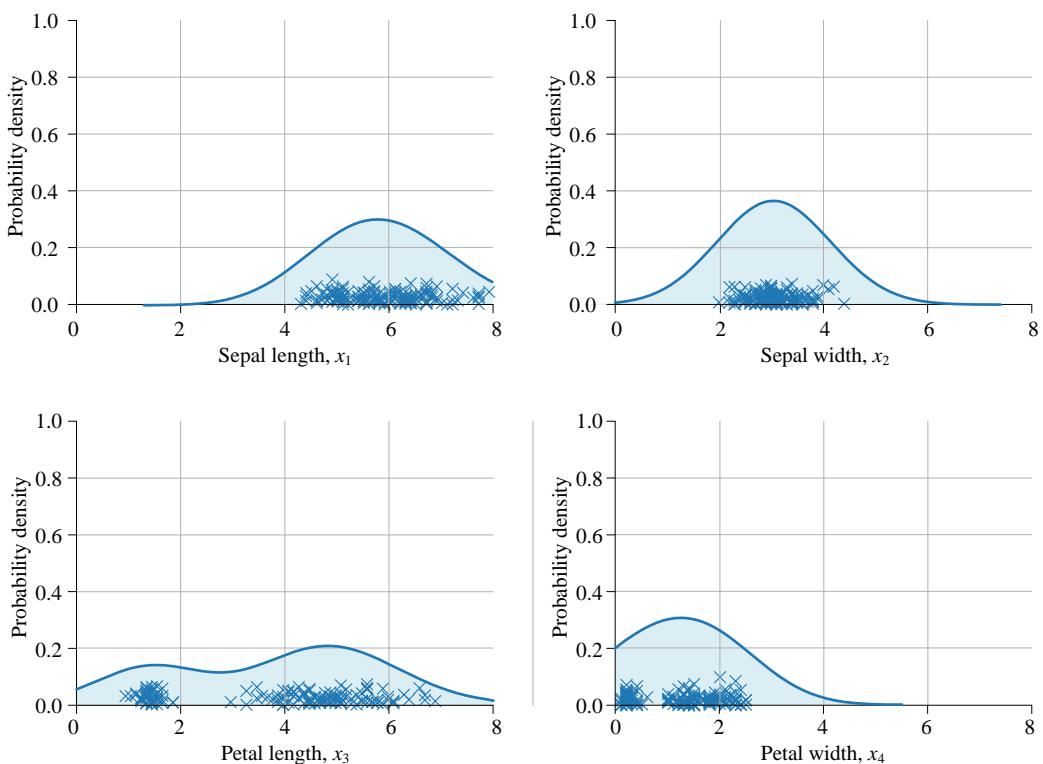


图 11. 鸢尾花四个特征数据的概率密度函数曲线, $h = 0.1$



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 12. 鸢尾花四个特征数据的概率密度函数曲线, $h = 1$

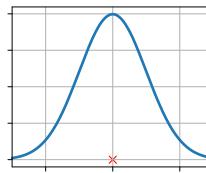
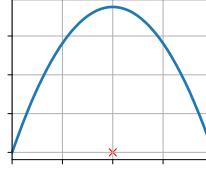
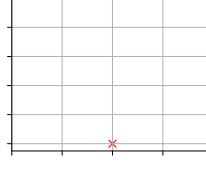
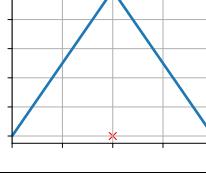
17.4 核函数：8 种常见核函数

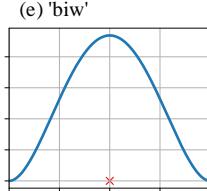
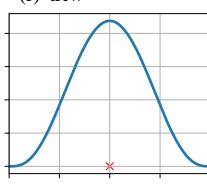
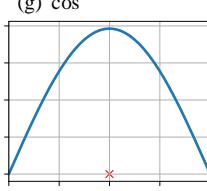
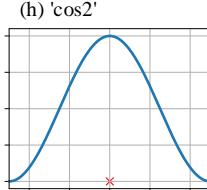
总结来说，核函数需要满足两个重要条件：(1) 对称性；(2) 面积为 1。用公式表达：

$$\begin{aligned} K(x) &= K(-x) \\ \int_{-\infty}^{+\infty} K(x) dx &= \frac{1}{h} \int_{-\infty}^{+\infty} K\left(\frac{x}{h}\right) dx = 1 \end{aligned} \quad (9)$$

表 1 总结 8 种满足以上两个条件的常用核函数。图 13 所示为这 8 种不同核函数估计得到的鸢尾花萼长度概率密度曲线。

表 1.8 种常见核函数

核函数	函数	函数图像
Gaussian	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$	(a) 'gau' 
Epanechnikov	$K(x) = \frac{3}{4}(1-x^2), x \leq 1$	(b) 'epa' 
Uniform	$K(x) = \frac{1}{2}, x \leq 1$	(c) 'uni' 
Triangular	$K(x) = 1- x , x \leq 1$	(d) 'tri' 

Biweight	$K(x) = \frac{15}{16}(1-x^2)^2, x \leq 1$	(e) 'biw' 
Triweight	$K(x) = \frac{35}{32}(1-x^2)^3, x \leq 1$	(f) 'triw' 
Cosine	$K(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}x\right), x \leq 1$	(g) 'cos' 
Cosine2	$K(x) = 1 + \cos(2\pi x), x \leq \frac{1}{2}$	(h) 'cos2' 

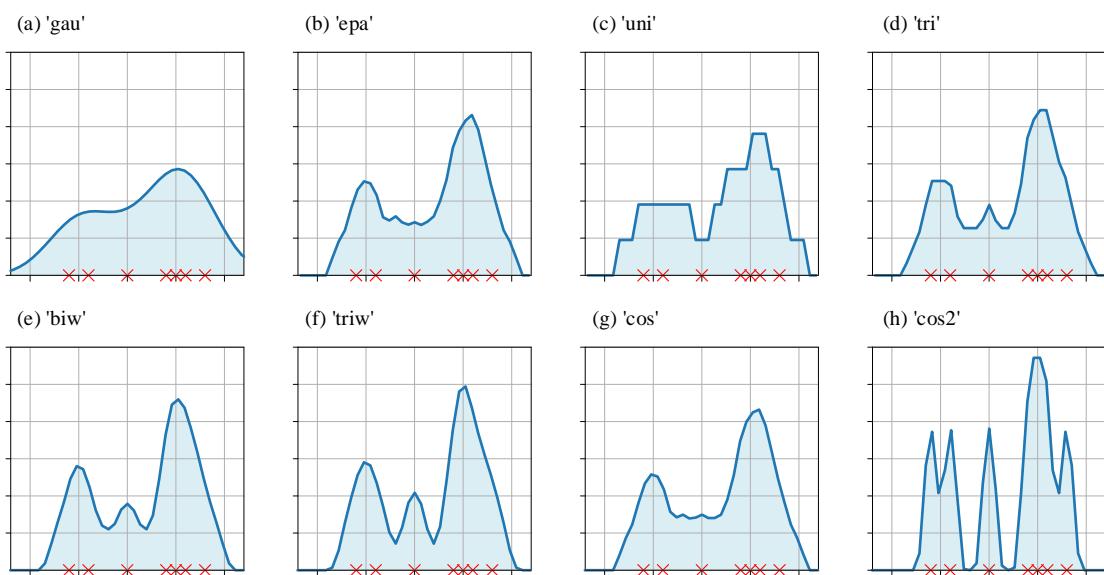
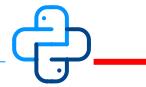


图 13. 八个不同核函数得到的不同的概率密度估计



Bk5_Ch17_03.py 代码绘制表 1 和图 13。也请大家学习使用 `sklearn.neighbors.KernelDensity()` 函数获得概率密度估计曲线。

17.5 二元 KDE：概率密度曲面

二元，乃至多元 KDE 的原理和前文所述的一元 KDE 完全相同。对于 n 个多维样本数据点 $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$ ，如下多个核函数叠加、再平均便得到概率密度估计：

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n K_H(\mathbf{x} - \mathbf{x}^{(i)}) \quad (10)$$

⚠ 注意， 默认 \mathbf{x} 和 $\mathbf{x}^{(i)}$ 均为列向量。 $\mathbf{x}^{(i)}$ 起到平移作用。

高斯核函数

高斯核函数 $K_H(\mathbf{x})$ 的定义为：

$$K_H(\mathbf{x}) = \det(\mathbf{H})^{-\frac{1}{2}} K\left(\mathbf{H}^{-\frac{1}{2}} \mathbf{x}\right) \quad (11)$$

带宽的形式为矩阵 \mathbf{H} ， \mathbf{H} 为正定矩阵。以二元高斯核函数为例， $K(\mathbf{x})$ 定义为：

$$K(\mathbf{x}) = \frac{1}{2\pi} \exp\left(-\frac{\mathbf{x}^T \mathbf{x}}{2}\right) \quad (12)$$

图 14 所示为高斯核二元 KDE 原理。图中，每个样本点都用一个 IID 二元高斯分布曲面描述。这些曲面先叠加、再平均便获得二元高斯核 KDE 估计得到概率密度曲面。

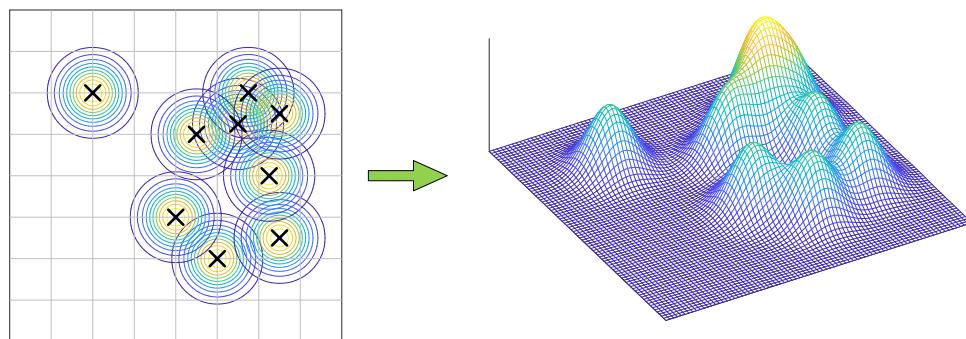


图 14. 二元高斯 KDE 原理

以鸢尾花数据为例

图 15 和图 16 所示为鸢尾花花萼长度和花萼宽度两个特征数据的 KDE 曲面。

`sklearn.neighbors.KernelDensity()` 函数也可以用来概率密度估计。注意，这个函数返回的是对数概率密度 $\ln(\text{PDF})$ 。

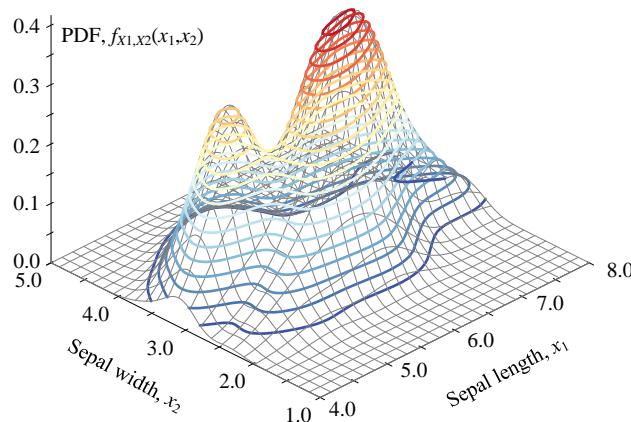


图 15. 鸢尾花花萼长度和花萼宽度两个特征数据的 KDE 曲面

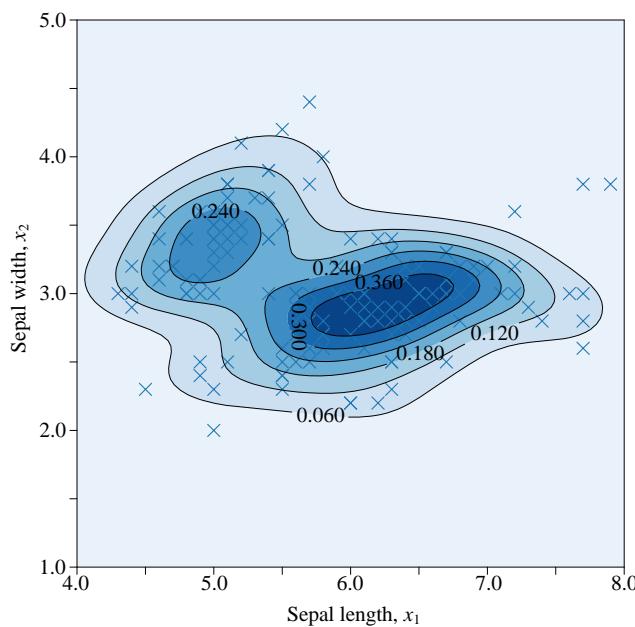
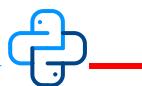


图 16. 鸢尾花花萼长度和花萼宽度两个特征数据的 KDE 曲面等高线图



Bk5_Ch17_04.py 代码绘制图 15 和图 16。Bk6_Ch17_05.py 用 Seaborn 绘制 KDE 曲面等高线。



在实际应用中，概率密度估计可以用来描述和模拟数据的分布特征，进行分类、聚类、异常检测等数据挖掘任务，也可以用于模型选择和参数估计。

常见的概率密度估计方法包括核密度估计、直方图估计、参数估计等。本章主要介绍的是核密度估计。核密度估计是一种非参数方法，可以用来估计连续随机变量的概率密度函数。请大家务必掌握高斯核密度估计，我们会在贝叶斯分类中看到这个工具的应用。

本书有关频率学派的内容到此结束。前文反复提过，《统计至简》重贝叶斯学派，轻频率学派，下面连续五章我们将看到贝叶斯定理在分类、推断两类问题的应用。

18

Bayesian Classification

贝叶斯分类

最大化后验概率，利用花萼长度分类鸢尾花



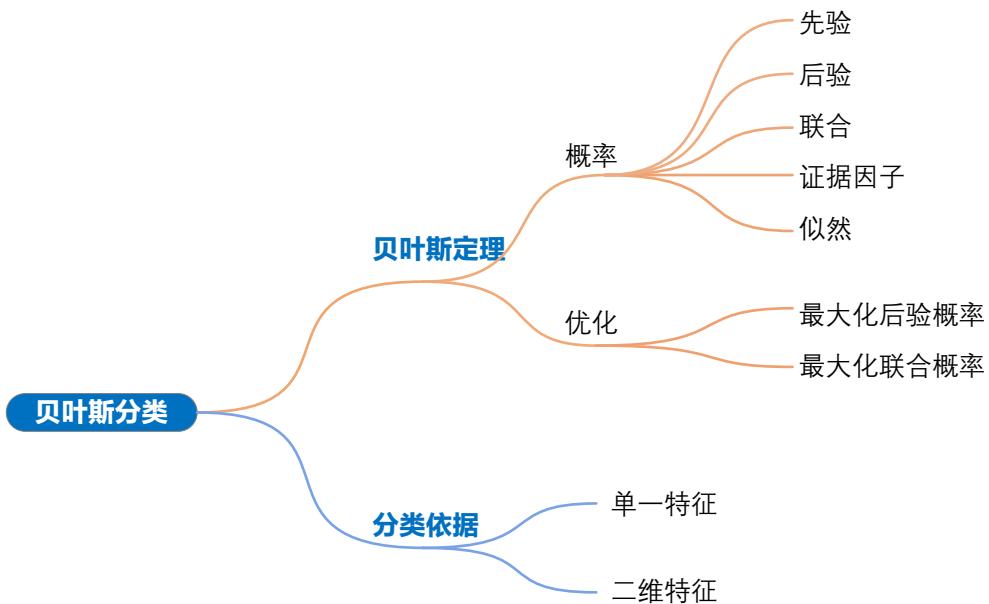
我们认为用最简单的假设来解释现象是一个很好的原则。

We consider it a good principle to explain the phenomena by the simplest hypothesis possible.

—— 托勒密 (Ptolemy) | 数学家、天文学家、地心说提出者 | 100 ~ 170



- ◀ matplotlib.pyplot.fill_between() 区域填充颜色
- ◀ seaborn.kdeplot() 绘制 KDE 概率密度估计曲线
- ◀ statsmodels.api.nonparametric.KDEUnivariate() 构造一元 KDE
- ◀ statsmodels.nonparametric.kde.kernel_switch() 更换核函数
- ◀ statsmodels.nonparametric.kernel_density.KDEMultivariate() 构造多元 KDE



18.1 贝叶斯定理：分类鸢尾花

本章和下一章和读者探讨采用贝叶斯定理对鸢尾花数据分类。本章采用鸢尾花数据中的花瓣长度作为研究对象，利用 KDE 生成概率密度函数，预测鸢尾花分类。

以下是使用贝叶斯定理进行分类的一般步骤：

- ▶ 收集数据，并提取特征。
- ▶ 对于每个类别，计算其在所有样本中出现的概率，称之为先验概率。
- ▶ 对于每个特征，计算它在每个类别下的概率，称之为条件概率。
- ▶ 根据贝叶斯定理，计算给定特征下，每个类别出现的概率，称之为后验概率。
- ▶ 根据后验概率的大小判定分类。

具体实现过程中，可以使用不同的算法来计算条件概率和后验概率，如朴素贝叶斯算法、高斯朴素贝叶斯算法等。同时，为了避免过拟合和欠拟合问题，我们还需要使用交叉验证、平滑等技术来提高分类器的性能。

为了帮助大家理解贝叶斯分类，我们首先回忆贝叶斯定理。

贝叶斯定理

大家知道鸢尾花数据分为三类——setosa、versicolour、virginica。我们分别用 C_1 、 C_2 、 C_3 作为标签代表这三类鸢尾花。

对于鸢尾花分类问题，贝叶斯定理可以按如下方式表达：

$$\underbrace{f_{Y|X}(C_k|x)}_{\text{Posterior}} = \frac{\overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}}}{\overbrace{f_X(x)}^{\text{Evidence}}} = \frac{\overbrace{f_{X|Y}(x|C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}}}{\overbrace{f_X(x)}^{\text{Evidence}}}, \quad k=1,2,3 \quad (1)$$

其中， X 代表鸢尾花花瓣长度的连续随机变量， Y 代表分类的离散随机变量， Y 的取值为 C_1 、 C_2 、 C_3 。

下面我们给 (1) 中几个概率值取名字：

$f_{Y|X}(C_k|x)$ 为**后验概率** (posterior)，又叫**成员值** (membership score)。在给定任意花瓣长度 x 的条件下，比较三个后验概率 $f_{Y|X}(C_1|x)$ 、 $f_{Y|X}(C_2|x)$ 、 $f_{Y|X}(C_3|x)$ 大小，可以作为判定鸢尾花分类的依据。

$f_{X,Y}(x, C_k)$ 为**联合概率** (joint)，也可以记做 $f_{X\cap Y}(x\cap C_k)$ 。

$f_X(x)$ 为**证据因子** (evidence), 也叫证据。证据因子和分类无关, 仅代表鸢尾花花萼长度 X 的概率分布情况。(1) 中, 证据因子 $f_X(x)$ 对联合概率 $f_{X,Y}(x,C_k)$ 进行**归一化** (normalization) 处理。本章假设 $f_X(x) > 0$ 。

$p_Y(C_k)$ 为**先验概率** (prior), 表达样本集合中 $C_k (k = 1, 2, 3)$ 类样本占比。注意, $p_Y(C_k)$ 为概率质量函数; 这是因为随机变量 Y 为离散随机变量, 取值为 $Y = C_1, C_2, C_3$ 。

$f_{Y|X}(x|C_k)$ 为**似然概率** (likelihood)。白话解释, 给定类别 C_k 中 x 出现的可能性, 比如给定鸢尾花为 setosa, 花萼长度为 10 cm 的可能性可以写成 $f_{Y|X}(10 | \text{Setosa})$ 。

图 1 可可视化三分类问题中的贝叶斯定理。下面, 我们逐一讲解上述不同的概率, 以及它们如何帮助我们完成鸢尾花分类。

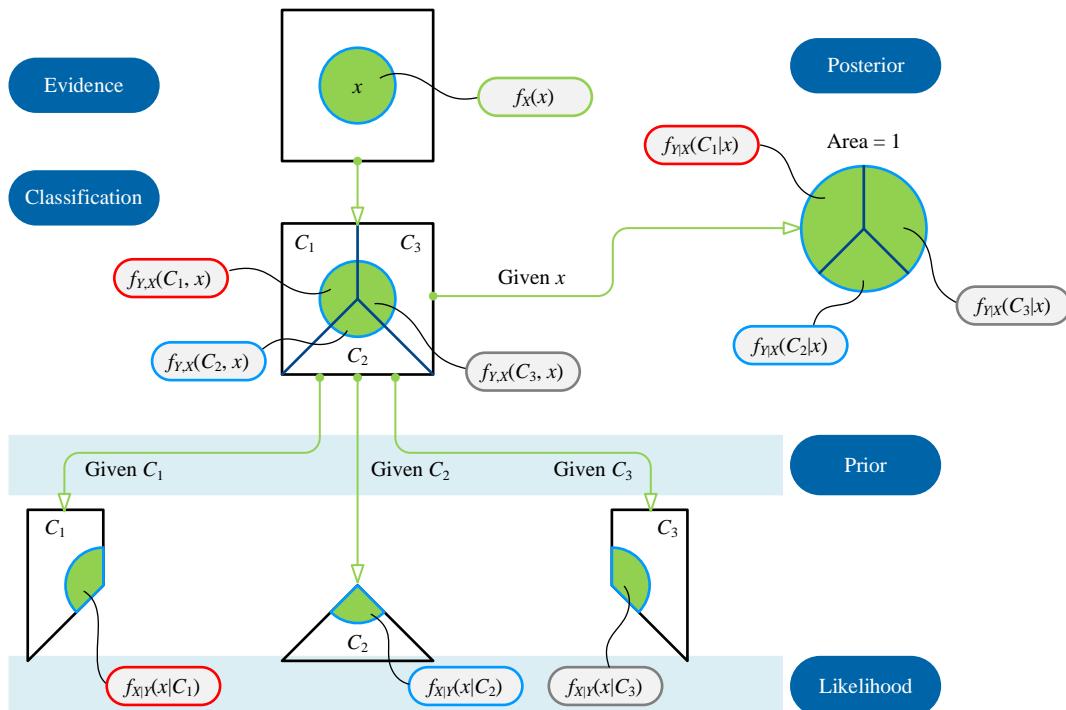


图 1. 利用贝叶斯定理, 以花萼长度作为特征对鸢尾花进行分类

18.2 似然概率：给定分类条件下的概率密度

似然概率 $f_{X|Y}(x|C_k)$ 本身是条件概率, 它描述的是给定类别 $Y = C_k$ 中 $X = x$ 出现的可能性。

注意, 本章中 $f_{X|Y}(x|C_k)$ 为概率密度函数 PDF。

图 2 (a)、(b)、(c) 分别展示 $f_{X|Y}(x|C_1)$ 、 $f_{X|Y}(x|C_2)$ 、 $f_{X|Y}(x|C_3)$ 三个似然概率 PDF 曲线。这三条概率密度曲线采用高斯 KDE 估计得到。

在鸢尾花数据集所有 150 个样本数据中如果，我们只分析标签为 C_1 (Setosa) 的 50 个样本的话， $f_{X|Y}(x|C_1)$ 就是这 50 个样本数据得到花萼长度的概率密度函数 PDF。

$f_{X|Y}(x|C_2)$ 代表给定鸢尾花分类为 C_2 (Versicolour)，花萼长度的概率密度函数。同理， $f_{X|Y}(x|C_3)$ 代表给定鸢尾花分类为 C_3 (Virginica)，花萼长度的概率密度函数。图 2 (c) 比较 $f_{X|Y}(x|C_1)$ 、 $f_{X|Y}(x|C_2)$ 、 $f_{X|Y}(x|C_3)$ 三条曲线。

⚠ 注意， $f_{X|Y}(x|C_k)$ 和横轴包裹的面积为 1。

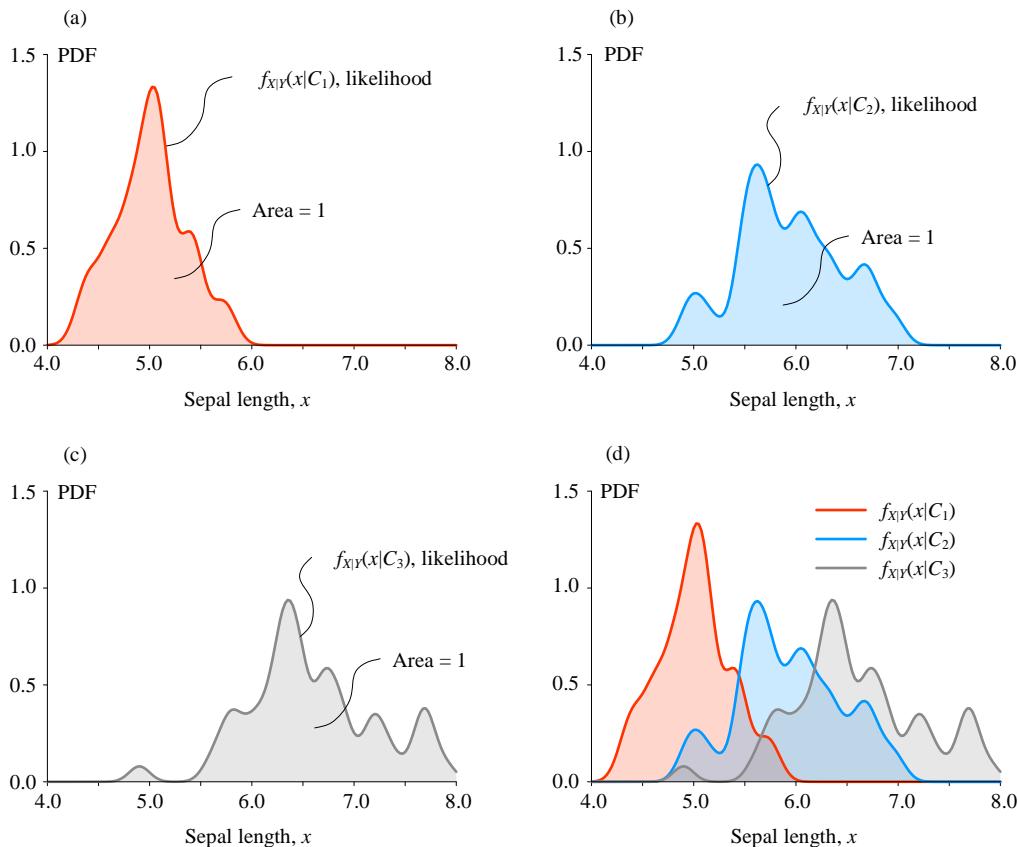


图 2. 三个似然概率 PDF 曲线 $f_{X|Y}(x|C_k)$

18.3 先验概率：鸢尾花分类占比

先验概率 $p_Y(C_k)$ 描述的是样本集合中 C_k 类样本占比。由于 Y 为离散随机变量，因此我们采用概率质量函数。 $p_Y(C_k)$ 具体计算如下：

$$p_Y(C_k) = \frac{\text{count}(C_k)}{\text{count}(\Omega)}, \quad k=1,2,3 \quad (2)$$

其中，`count()` 为计数运算符，`count(C_k)` 计算标签样本空间 Ω 中 C_k 类样本数据数量。

如图 3 所示，对于鸢尾花数据，每一类标签的样本数据都是 50，因此三类标签的先验概率都是 $1/3$ ：

$$p_Y(C_k) = \frac{50}{150} = \frac{1}{3}, \quad k=1,2,3 \quad (3)$$

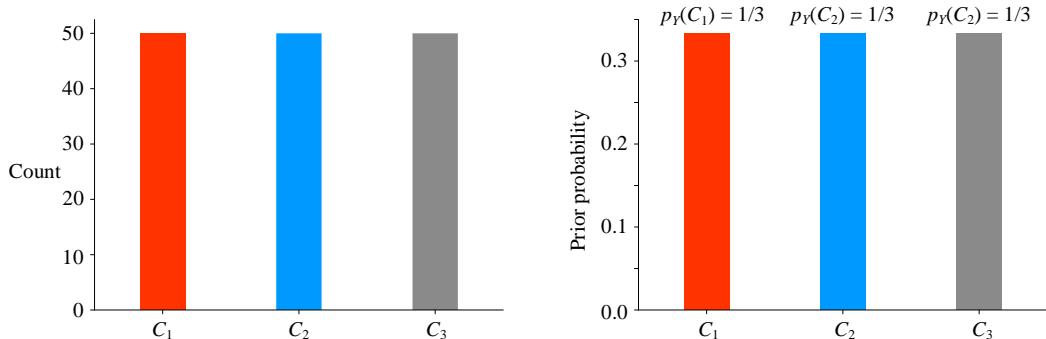


图 3. 150 个样本数据总三类的频数和先验概率

18.4 联合概率：可以作为分类标准

联合概率 $f_{X,Y}(x, C_k)$ 描述事件 $Y = C_k$ 和事件 $X = x$ 同时发生的可能性。

比如，花萼长度为 $x = 5.6$ cm 且鸢尾花分类为 $Y = C_1$ (Setosa) 的可能性可以用 $f_{X,Y}(5.6, C_1)$ 表达。

⚠ 注意， $f_{X,Y}(x, C_k)$ 也是概率密度函数 PDF，并不是“概率”。

根据贝叶斯定理，联合概率 $f_{X,Y}(x, C_k)$ 可以通过似然概率 $f_{X|Y}(x|C_k)$ 和先验概率 $p_Y(C_k)$ 相乘得到：

$$\overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}} = \overbrace{f_{X|Y}(x|C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}} \quad (4)$$

图 4 (a)、(b)、(c) 分别展示 $f_{X,Y}(x, C_1)$ 、 $f_{X,Y}(x, C_2)$ 、 $f_{X,Y}(x, C_3)$ 三个联合概率 PDF 曲线。这三幅图还展示了从似然概率 $f_{X|Y}(x|C_k)$ 到联合概率 $f_{X,Y}(x, C_k)$ 的缩放过程。

似然概率 $f_{X|Y}(x|C_k)$ 和横轴包裹的面积为 1。而联合概率 $f_{X,Y}(x, C_k)$ 和横轴包裹的面积为 $p_Y(C_k)$ 。

图 4 (d) 比较 $f_{X,Y}(x, C_1)$ 、 $f_{X,Y}(x, C_2)$ 、 $f_{X,Y}(x, C_3)$ 三个联合概率 PDF 曲线，即“似然概率 \times 先验概率”。实际上，这三条曲线的高低已经可以用来作为分类标准，这是本章后续要介绍的内容。

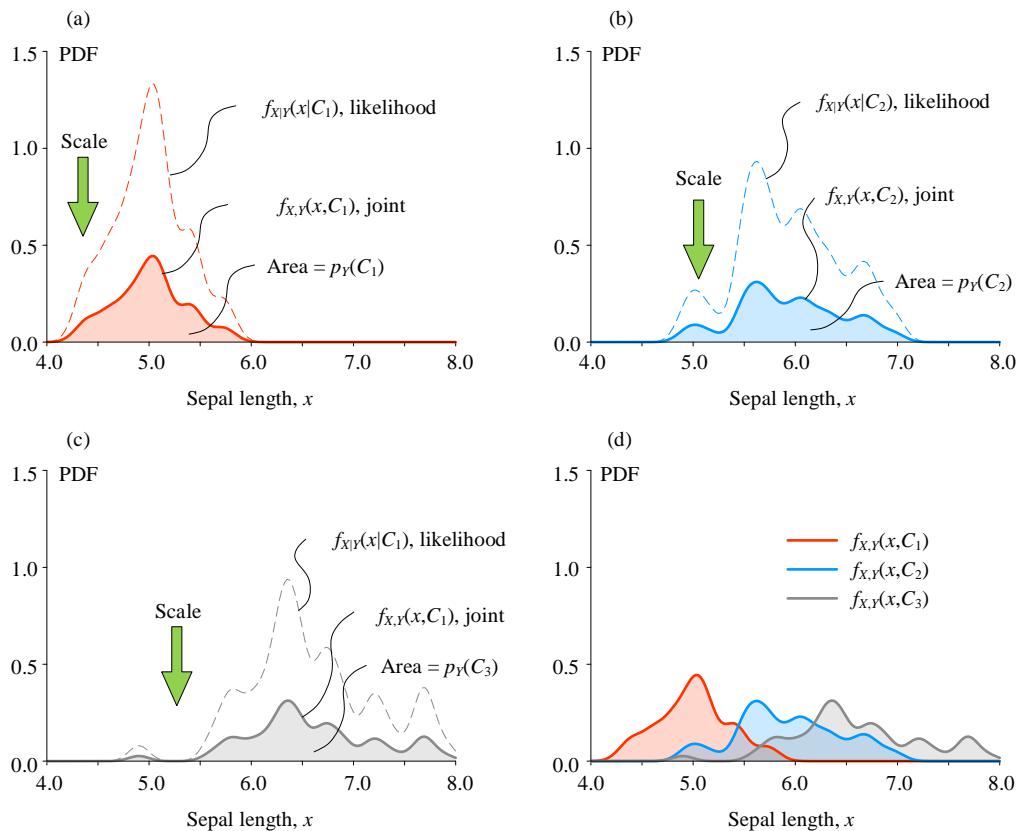


图 4. 先验概率和联合概率的关系

18.5 证据因子：和分类无关

证据因子 $f_X(x)$ 实际上就是 X 的边缘概率密度函数 PDF，证据因子和分类无关。对于本章鸢尾花数据， $f_X(x)$ 就是根据样本数据利用 KDE 方法估计得到的概率密度函数。

显然，对于鸢尾花样本数据， C_1 、 C_2 、 C_3 为一组不相容分类，对样本空间 Ω 形成分割。根据全概率定理，下式成立：

$$\overbrace{f_X(x)}^{\text{Evidence}} = \sum_{k=1}^3 \overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}} = \sum_{k=1}^3 \overbrace{f_{X|Y}(x|C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}} \quad (5)$$

也就是说，似然概率密度 $f_{X|Y}(x|C_k)$ 和先验概率 $p_Y(C_k)$ ，可以用来估算 $f_X(x)$ 。

对于鸢尾花三分类，(5) 可以展开来写：

$$f_X(x) = f_{X,Y}(x, C_1) + f_{X,Y}(x, C_2) + f_{X,Y}(x, C_3) \quad (6)$$

图 5 所示为利用联合概率 PDF 计算证据因子 PDF 的过程。

⚠ 注意， $f_X(x)$ 和横轴包裹的面积为 1。

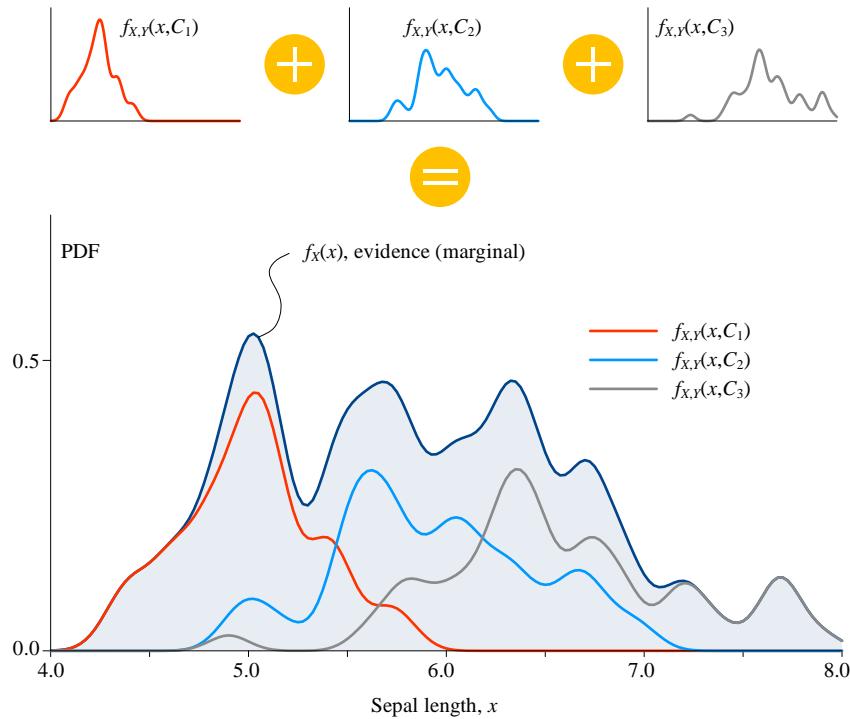


图 5. 叠加联合概率曲线，估算证据因子概率密度函数

18.6 后验概率：也是分类的依据

$f_{Y|X}(C_k | x)$ 指的是在事件 $X = x$ 发生条件下，事件 $Y = C_k$ 发生的概率。后验概率 $f_{Y|X}(C_k | x)$ 又叫成员值 (membership score)。

白话来说，后验概率指的是在已知一些先验条件的情况下，通过贝叶斯定理计算得出的条件概率。换句话说，它是指在观测到某些数据或证据后，对于假设的某个事件发生的概率的更新。

比如，给定花萼的长度为 $x = 5.6 \text{ cm}$ ，鸢尾花被分类为 $Y = C_1$ (Setosa) 的可能性，就可以用 $f_{Y|X}(C_1 | 5.6)$ 来描述。

⚠ 注意，后验概率实际上是概率，不是概率密度。因此， $f_{Y|X}(C_k | x)$ 的取值范围为 $[0, 1]$ 。

根据贝叶斯定理，当 $f_X(x) > 0$ 时，后验概率 PDF $f_{Y|X}(C_k | x)$ 可以根据下式计算得到：

$$\overbrace{f_{Y|X}(C_k | x)}^{\text{Posterior}} = \frac{\overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}}}{\underbrace{f_X(x)}_{\text{Evidence}}} \quad (7)$$

图 6 所示为后验概率 PDF 曲线 $f_{Y|X}(C_1 | x)$ 的计算过程。图 7 则比较另外两组联合概率、证据因子、后验概率曲线。

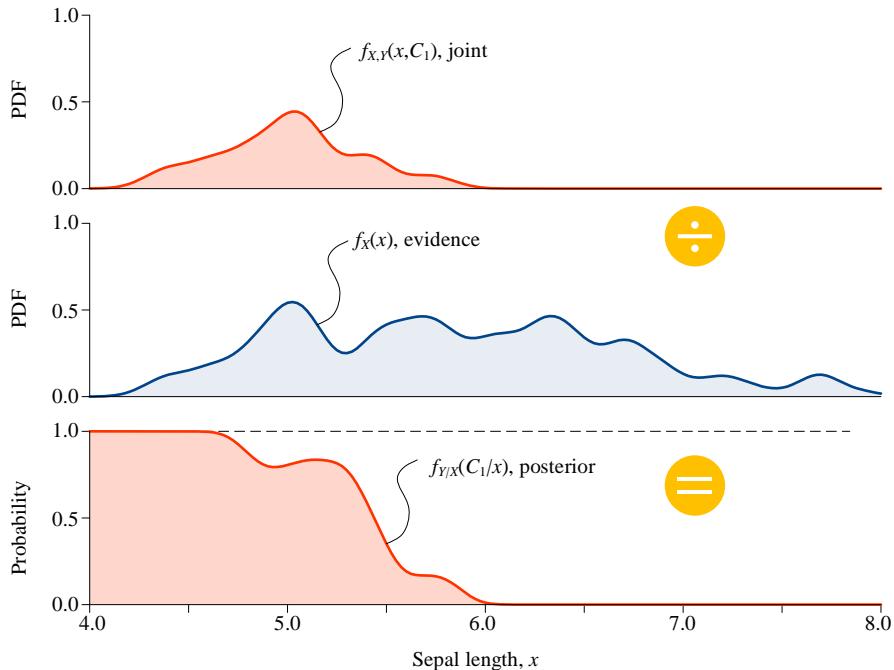
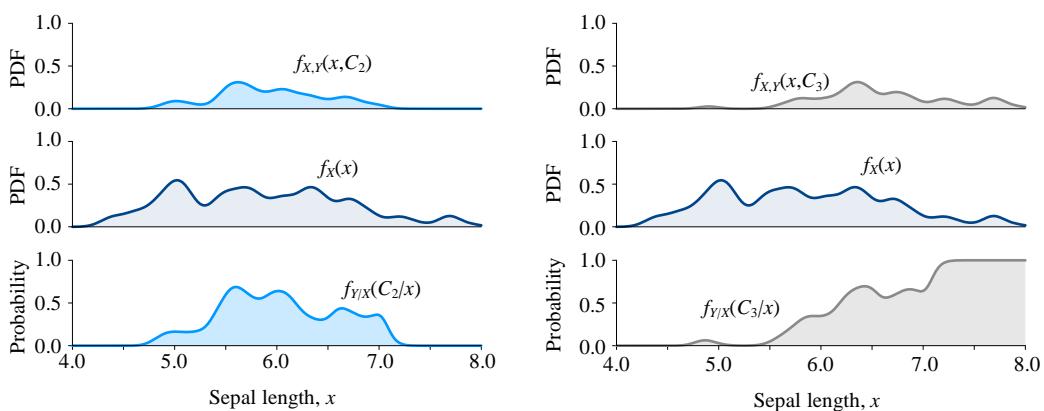
图 6. 计算后验概率 PDF 曲线 $f_{y|x}(C_1|x)$ 

图 7. 比较联合概率、证据因子、后验概率曲线

成员值

后验概率之所以被称作“成员值”是因为：

$$\sum_{k=1}^3 \underbrace{f_{y|x}(C_k | x)}_{\text{Posterior}} = 1 \quad (8)$$

这个式子不难推导。根据贝叶斯定理，下式成立：

$$\overbrace{f_X(x)}^{\text{Evidence}} = \sum_{k=1}^3 \overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}} = \sum_{k=1}^3 \overbrace{f_{Y|X}(C_k|x)}^{\text{Posterior}} \overbrace{f_X(x)}^{\text{Evidence}} \quad (9)$$

即，

$$\overbrace{f_X(x)}^{\text{Evidence}} = \overbrace{f_X(x)}^{\text{Evidence}} \sum_{k=1}^3 \overbrace{f_{Y|X}(C_k|x)}^{\text{Posterior}} \quad (10)$$

$f_X(x) > 0$ 时，(10) 左右消去 $f_X(x)$ 便获得 (8)。

分类依据

在给定任意花瓣长度 x 的条件下，比较三个后验概率 $f_{Y|X}(C_1|x)$ 、 $f_{Y|X}(C_2|x)$ 、 $f_{Y|X}(C_3|x)$ 大小，最大后验概率对应的标签就可以作为鸢尾花分类依据。

举个例子，某朵鸢尾花花瓣长度为 $x = 5.6$ cm 的前提下，它一定被分类为 C_1 、 C_2 、 C_3 任一标签。三种不同情况的可能性相加为 1，也就是说，这朵鸢尾花要么是 C_1 ，或者是 C_2 ，不然就是 C_3 。

换个角度来看，比较图 8 三条不同颜色曲线的高度，我们就可以据此判断鸢尾花的分类。

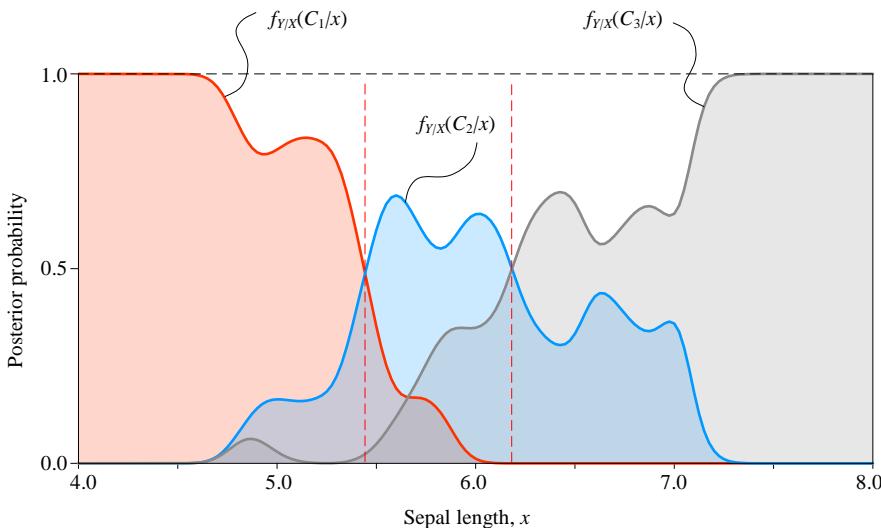


图 8. 比较三个后验概率 PDF 曲线 $f_{Y|X}(C_1|x)$ 、 $f_{Y|X}(C_2|x)$ 、 $f_{Y|X}(C_3|x)$

观察 (7)，可以发现后验概率 $f_{Y|X}(C_1|x)$ 正比于联合概率 $f_{X,Y}(x, C_k)$ ，证据因子 $f_X(x)$ 仅仅起到缩放作用：

$$\overbrace{f_{Y|X}(C_k|x)}^{\text{Posterior}} \propto \overbrace{f_{X,Y}(x, C_k)}^{\text{Joint}} = \overbrace{f_{X|Y}(x|C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}} \quad (11)$$

实际上，没有必要计算后验概率 $f_{Y|X}(C_1|x)$ ，比较联合概率 $f_{X,Y}(x, C_k)$ 就可以对鸢尾花进行分类。上式实际上是贝叶斯推断中最重要的正比关系——后验 \propto 似然 \times 先验。

比较四条曲线

本节最后，我们把**似然概率** (likelihood)、**联合概率** (joint)、**证据因子** (evidence)、**后验概率** (posterior) 这四条曲线放在一幅中加以比较，具体如图 9、图 10、图 11 所示。

请大家注意以下几点：

- ◀ 似然概率 (likelihood) 曲线为条件概率密度，和横轴围成图形的面积为 1；
- ◀ 似然概率 (likelihood) 经过先验概率 (prior) 缩放得到联合概率 (joint)；
- ◀ 后验 \propto 似然 \times 先验；
- ◀ 联合概率曲线面积为对应先验概率；
- ◀ 联合概率叠加得到证据因子 (evidence)；
- ◀ 联合概率 (joint) 除以证据因子得到后验概率 (posterior)，证据因子起到归一化作用；
- ◀ 后验概率，也叫成员值 (membership score)，本质上是概率值，取值范围在 [0, 1] 之间；
- ◀ 比较后验概率/成员值大小，可以预测分类，方便可视化。
- ◀ 比较联合概率密度 (似然 \times 先验) 大小，可以预测分类；

⚠ 再次强调，虽然放在同一张图上，图 9、图 10、图 11 中后验概率为具体概率值，而其他曲线均为概率密度函数。

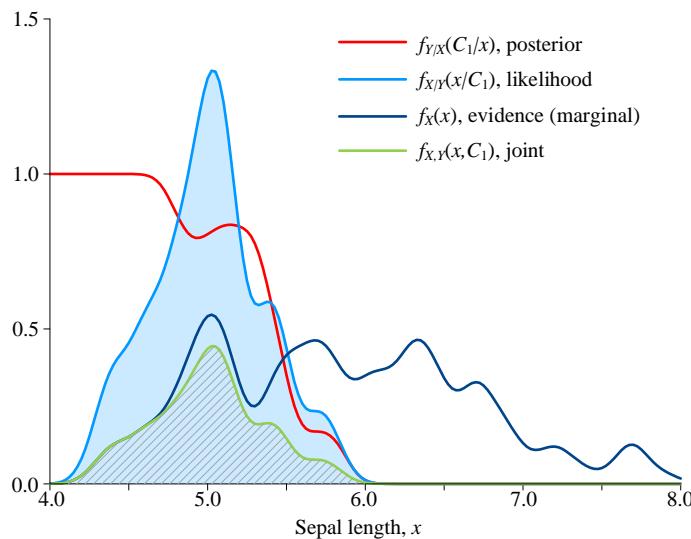
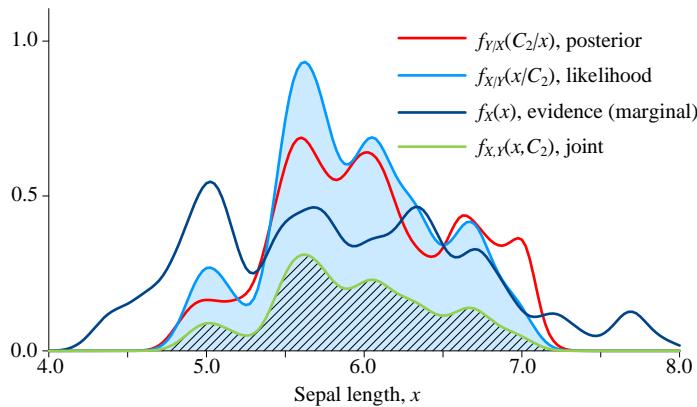
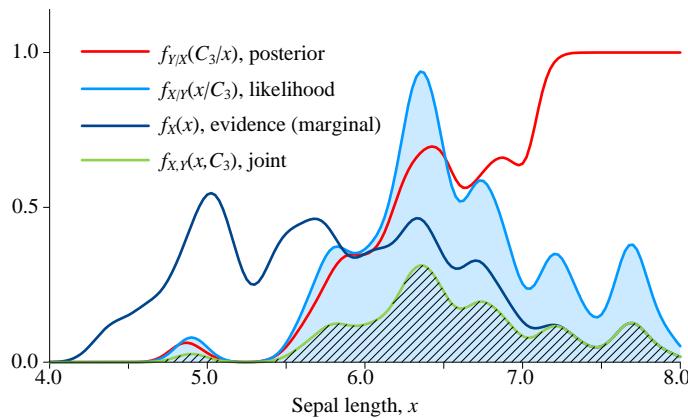
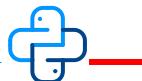


图 9. 比较后验概率 $f_{Y|X}(C_1 | x)$ 、似然概率 $f_{X|Y}(x|C_1)$ 、证据因子 $f_X(x)$ 、联合概率 $f_{X,Y}(x,C_1)$

图 10. 比较后验概率 $f_{Y|X}(C_2 | x)$ 、似然概率 $f_{X|Y}(x|C_2)$ 、证据因子 $f_X(x)$ 、联合概率 $f_{X,Y}(x,C_2)$ 图 11. 比较后验概率 $f_{Y|X}(C_3 | x)$ 、似然概率 $f_{X|Y}(x|C_3)$ 、证据因子 $f_X(x)$ 、联合概率 $f_{X,Y}(x,C_3)$ 

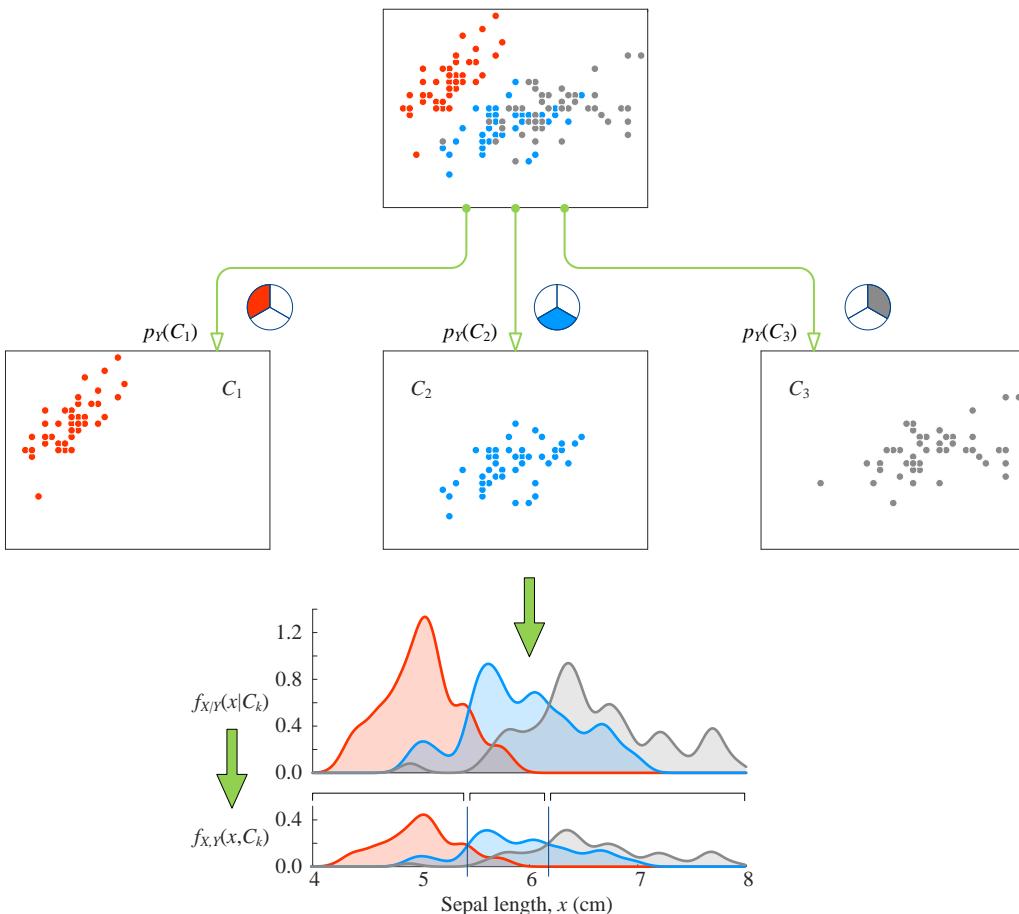
Bk5_Ch18_01.py 代码绘制本章前文大部分图像。

18.7 单一特征分类：基于 KDE

似然概率 → 联合概率

图 12 总结以花萼长度为单一特征，计算似然概率和联合概率的过程。

鸢尾花数据较为特殊，前文介绍过，鸢尾花数据共有 150 个数据点， C_1 、 C_2 和 C_3 三类各占 50，因此三个先验概率相等。因此，图 12 中，从似然概率密度 $f_{X|Y}(x | C_k)$ 到联合概率 $f_{X,Y}(x, C_k)$ ，高度缩放比例相同。一般情况下，相同缩放比例这种情况几乎不存在。

图 12. 似然概率到联合概率，花萼长度特征 x ，基于 KDE

比较后验概率

有了本节前文联合概率和证据因子，我们可以获得后验概率密度曲线，如图 13。后验概率也叫成员值，后验概率更容易分类可视化。

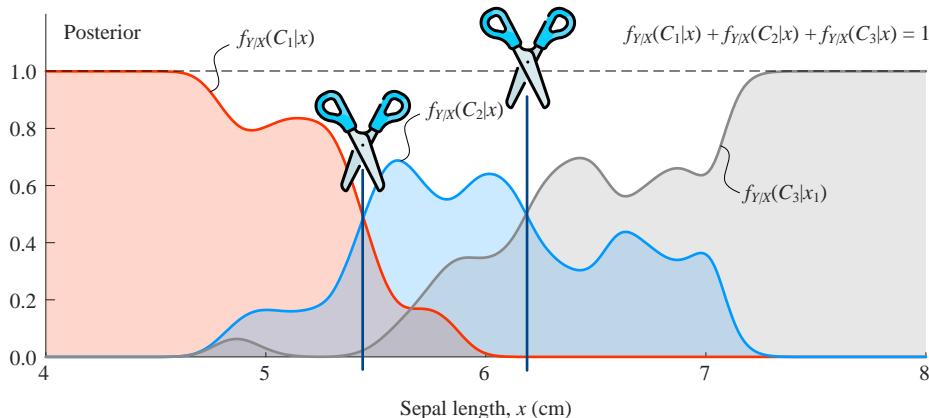


图 13. 后验概率，花萼长度特征，基于 KDE

举个例子

如图 14 所示，比较花萼长度特征后验概率大小，可以很容易预测 A 、 B 、 C 、 D 和 E 五点分类。 A 的预测分类为 C_1 ； B 为决策边界； C 的预测分类为 C_2 ； D 为**决策边界** (decision boundary)； E 预测分类为 C_3 。

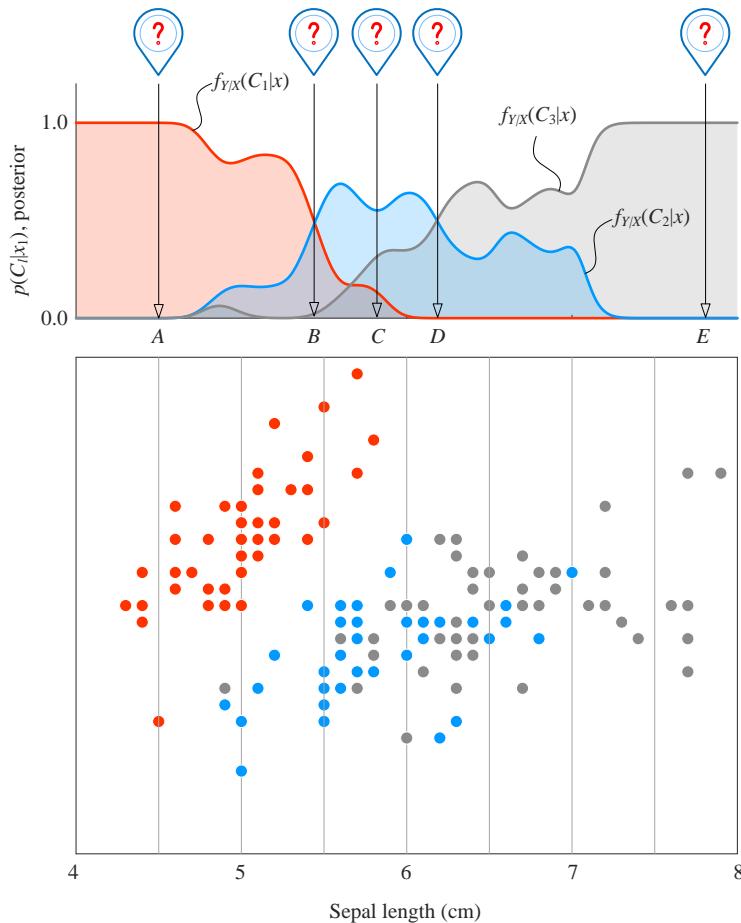


图 14. 利用花萼长度特征后验概率，进行分类预测

堆积直方图、饼图

图 15 所示为另外两种成员值（后验概率）的可视化方案——**堆积直方图** (stacked bar chart) 和 **饼图** (pie chart)。通过这两个可视化方案，大家可以清楚看到不同类别成员值随特征变化。

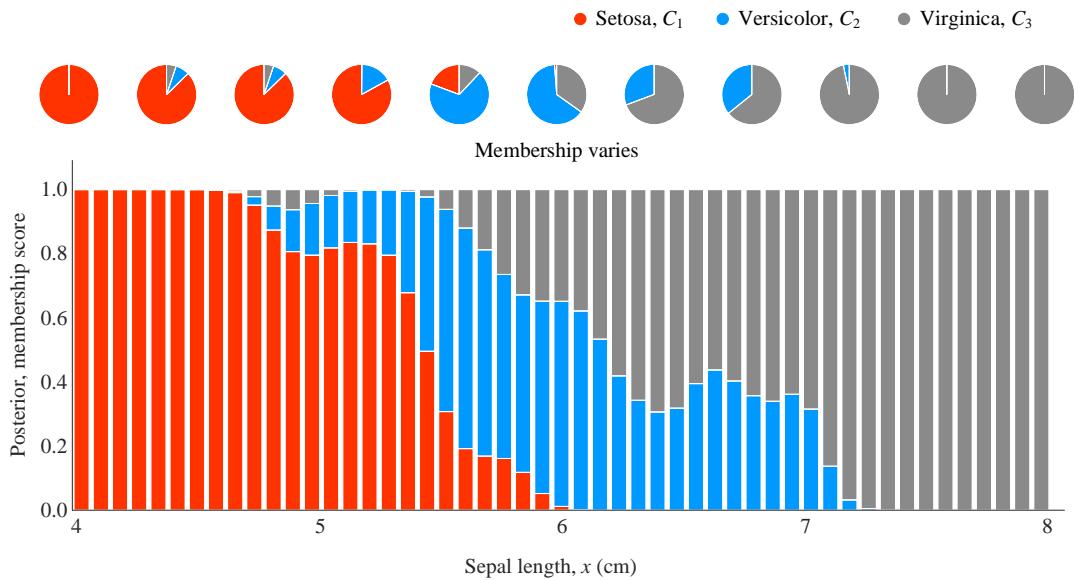


图 15. 堆积直方图和饼图，利用花萼长度特征成员值确定分类，基于 KDE

花萼宽度

本章前文都是基于花萼长度这个单一特征来判断鸢尾花分类，我们当然可以使用鸢尾花其他特征判断其分类。本节最后展示利用鸢尾花宽度作为依据判断鸢尾花分类。

图 16 所示为对于花萼宽度特征，从似然概率到联合概率的计算过程。

同理，比较花萼宽度特征的后验概率大小，可以决定图 17 中 A、B、C 和 D 点分类预测。A 的预测分类为 C₁；B 为决策边界；C 为决策边界；D 的预测分类为 C₂。

图 18 所示为利用花萼宽度特征成员值堆积直方图和饼图。

大家可能会问，如何同时利用鸢尾花花萼长度、花萼宽度作为分类依据？这个问题，我们下一章回答。

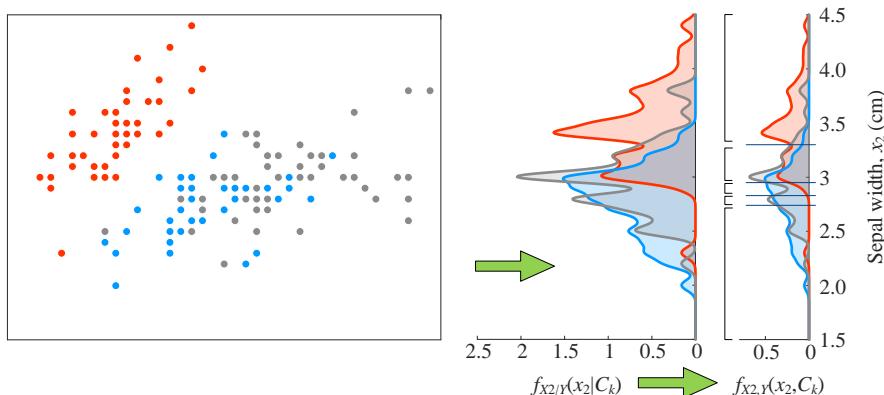


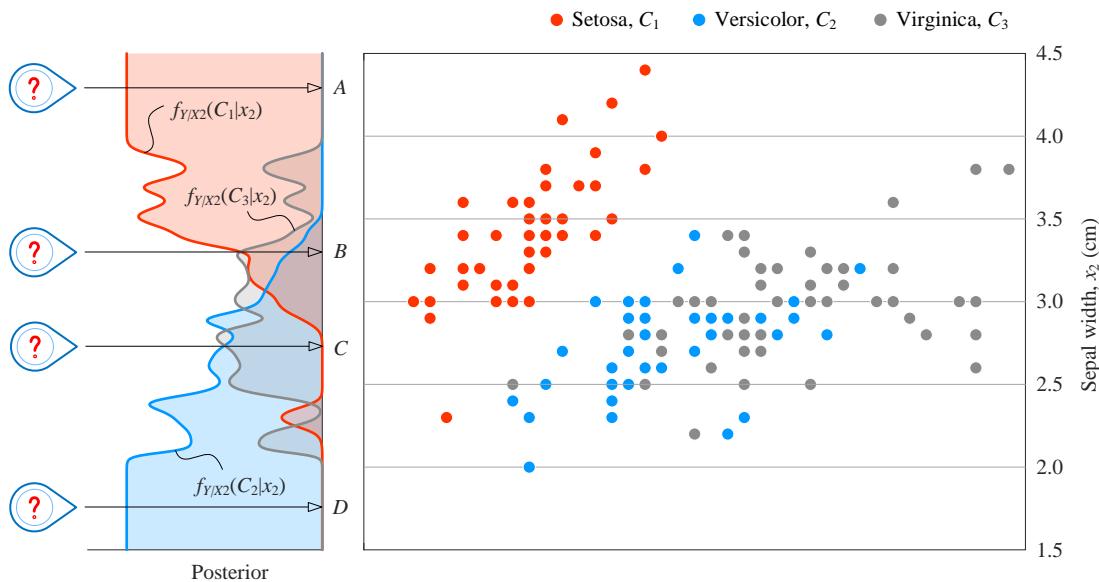
图 16. 似然概率到联合概率，花萼宽度特征 x_2 ，基于 KDE

图 17. 利用花萼宽度后验概率，进行分类预测

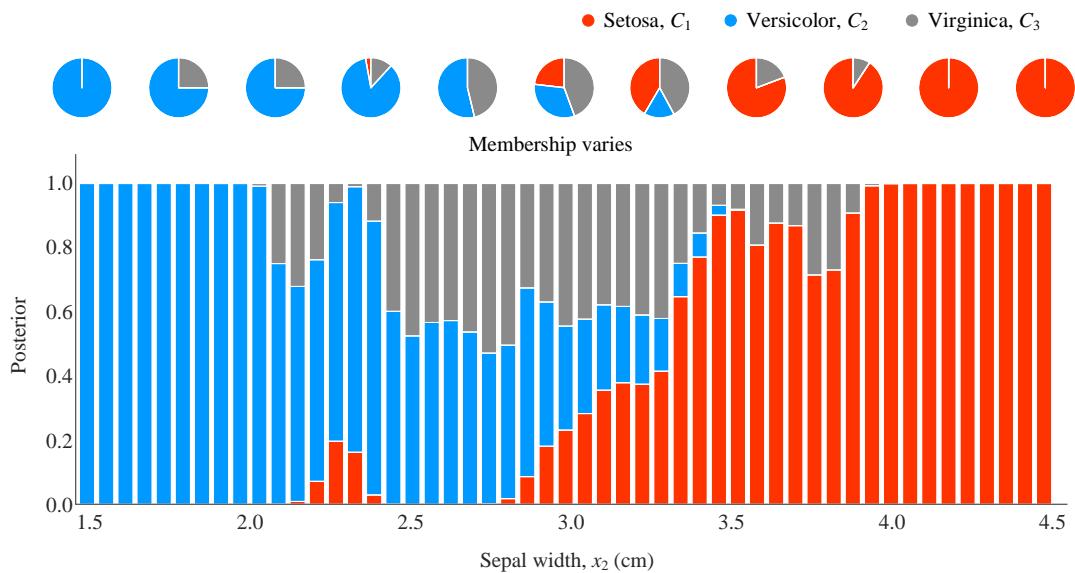


图 18. 堆积直方图和饼图，利用花萼宽度特征成员值确定分类，基于 KDE

18.8 单一特征分类：基于高斯

本章前文利用 KDE 方法估计似然概率，本章最后一节利用高斯分布估计似然概率。这一节，我们还是单独研究花萼长度特征 x_1 、花萼宽度特征 x_2 。

似然概率 → 联合概率

图 19 所示为花萼长度特征 x_1 上，利用一元高斯分布估算似然概率，然后计算联合概率；最后获得以特征 x_1 为依据决策边界。比较图 19 联合概率曲线高度，鸢尾花数据被划分为三个区域。这三个区域的位置和本章前文基于 KDE 估算稍有不同。

图 20 所示为花萼宽度特征 x_2 上同样过程。比较图 20 联合概率曲线高度，同样发现鸢尾花数据被划分为三个区域。

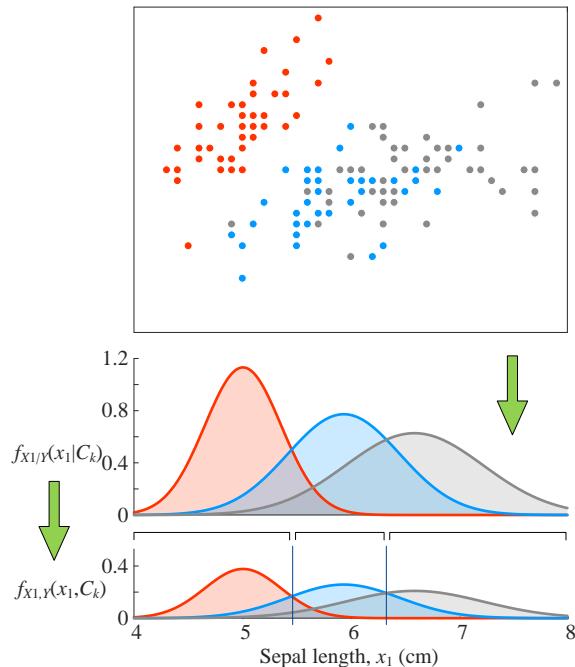


图 19. 似然概率到联合概率，花萼长度特征 x_1 ，基于高斯分布

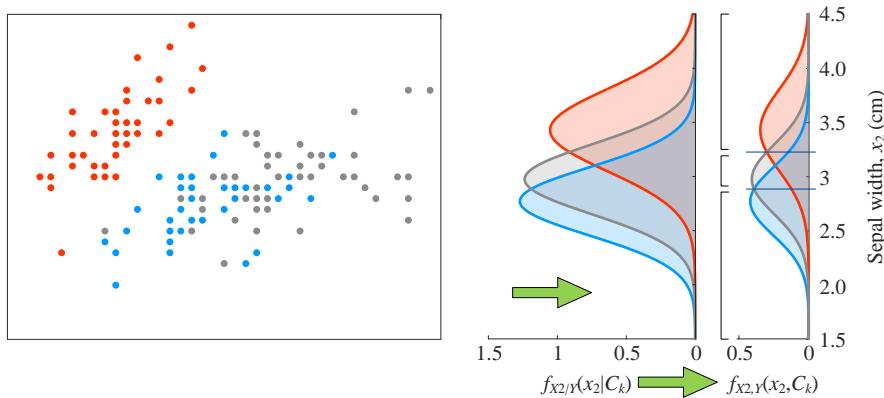


图 20. 似然概率到联合概率，花萼宽度特征 x_2 ，基于高斯分布

证据因子

图 21 和图 22 所示为利用全概率定理，获得 $f(x_1)$ 和 $f(x_2)$ 两个证据因子的概率密度函数。这实际上也是一种概率密度估算的方法。

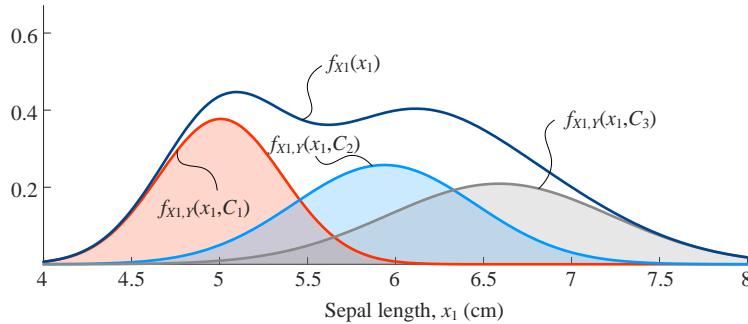


图 21. 证据因子/边缘概率，花萼长度特征 x_1 ，基于高斯分布

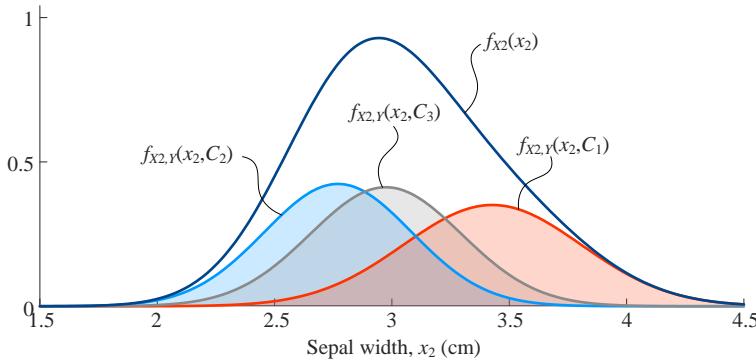


图 22. 证据因子/边缘概率，花萼宽度特征 x_2 ，基于高斯分布

后验概率

图 23 和图 24 比较两组后验概率曲线，以及如何据此得到的决策边界。

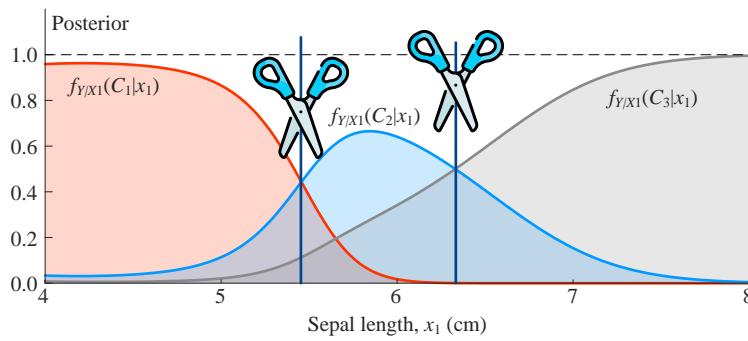
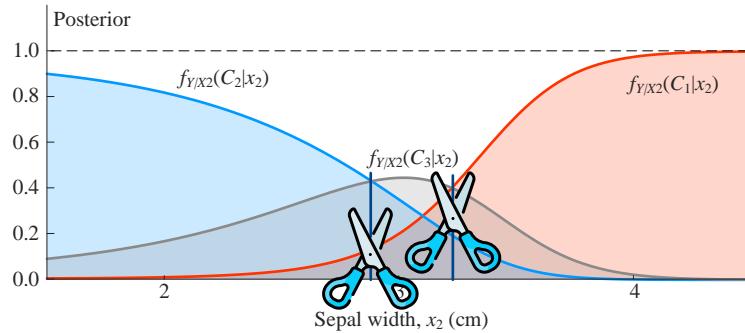
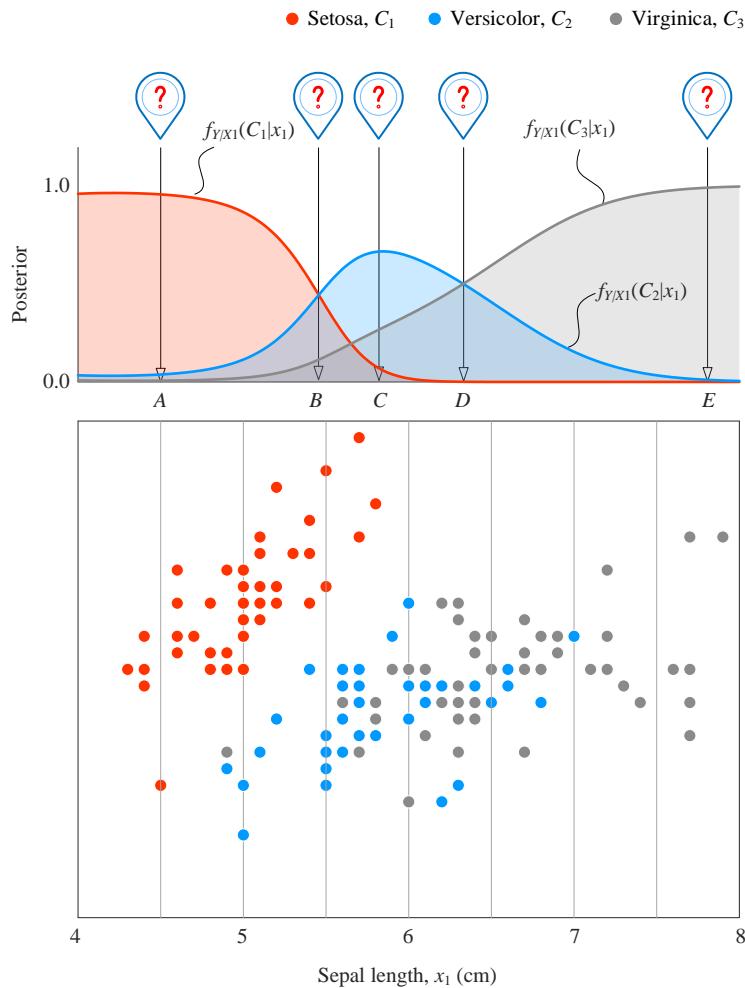


图 23. 后验概率，花萼长度特征 x_1 ，基于高斯分布

图 24. 后验概率，花萼宽度特征 x_2 ，基于高斯分布

后验概率：分类预测

图 25 所示为利用花萼长度特征后验概率曲线，进行分类预测。比较后验概率值大小可以判断：A 点预测分类为 C_1 ；B 点为 C_1 和 C_2 之间决策边界；C 点预测分类为 C_2 ；D 点为 C_2 和 C_3 之间决策边界；E 点预测分类为 C_3 。



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 25. 利用花萼长度特征后验概率，进行分类预测

图 26 所示为利用花萼宽度特征后验概率曲线，进行分类预测。比较后验概率值大小可以判断： A 点预测分类为 C_1 ； B 点预测分类为 C_3 ； C 点为 C_2 和 C_3 之间决策边界； D 点预测分类为 C_2 。

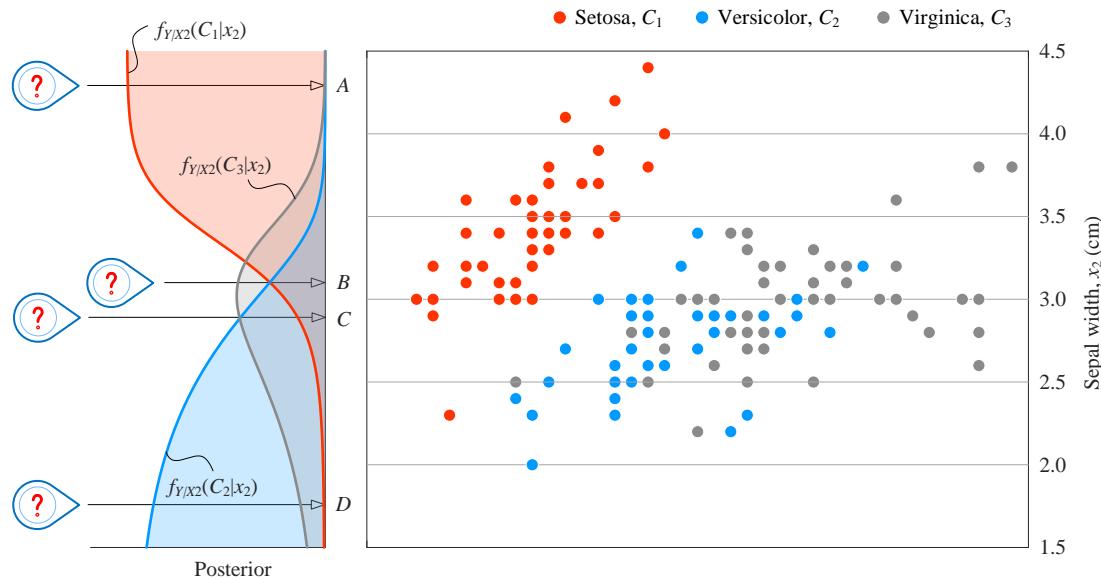


图 26. 利用花萼宽度特征后验概率，进行分类预测

图 27 和图 28 所示为利用堆积直方图和饼图表表达成员值/后验概率随特征变化。对比图 15 和图 18，可以发现，基于高斯分类的成员值/后验概率变化过程更为平滑。

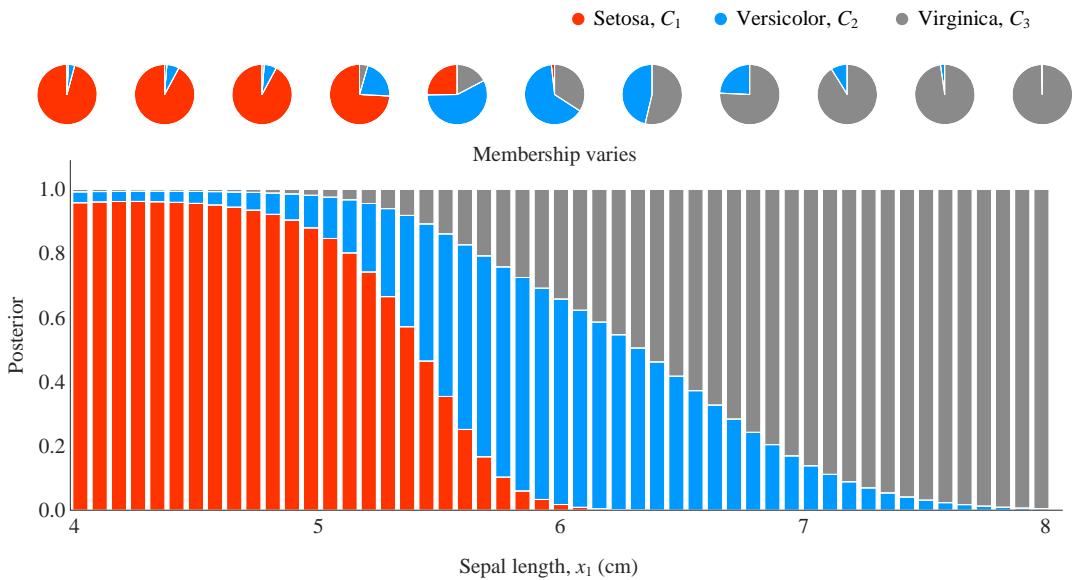


图 27. 堆积直方图和饼图，利用花萼长度特征成员值确定分类，基于高斯分布

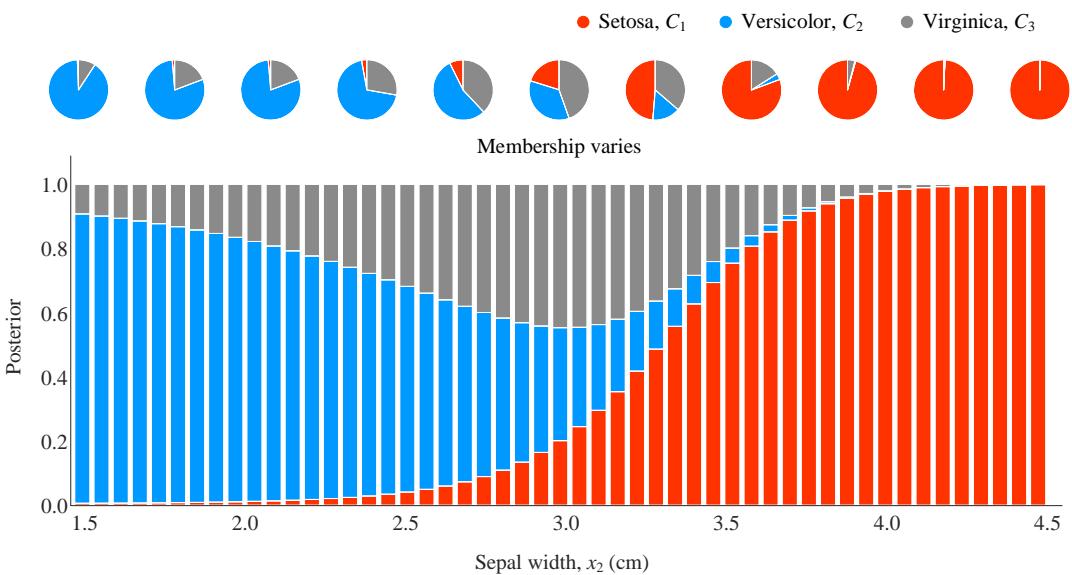
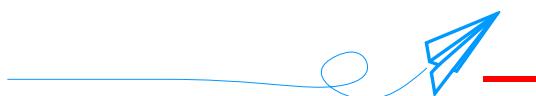


图 28. 堆积直方图和饼图，利用花萼宽度特征成员值确定分类，基于高斯分布



这一章中，大家必须要掌握的是贝叶斯定理中的先验概率、后验概率、证据因子、似然概率等概念。而贝叶斯分类是一种基于贝叶斯定理的分类方法。请大家务必掌握比例关系——后验 \propto 似然 \times 先验。这是贝叶斯推断中最重要的比例关系。

在贝叶斯分类算法中，优化问题可以最大化后验概率，也可以最大化联合概率，即“似然 \times 先验”。

下一章，我们将分类的依据从单一特征提高到二维，让大家更清楚地看到先验概率、后验概率、证据因子、似然概率的“样子”。下一章和本章内容安排几乎一致，可以对照阅读。

19

Dive into Bayesian Classification

贝叶斯分类进阶

计算后验概率，利用花萼长度和宽度分类鸢尾花



杀不死你的，会让你更强大。

What doesn't kill you, makes you stronger.

—— 弗里德里希·尼采 (Friedrich Nietzsche) | 德国哲学家 | 1844 ~ 1900



- ◀ `matplotlib.pyplot.contour3D()` 绘制三维等高线图
- ◀ `matplotlib.pyplot.contourf()` 绘制平面填充等高线
- ◀ `matplotlib.pyplot.fill_between()` 区域填充颜色
- ◀ `matplotlib.pyplot.plot_wireframe()` 绘制线框图
- ◀ `matplotlib.pyplot.scatter()` 绘制散点图
- ◀ `numpy.ones_like()` 用来生成和输入矩阵形状相同的全 1 矩阵
- ◀ `numpy.outer()` 计算外积，张量积
- ◀ `numpy.vstack()` 返回竖直堆叠后的数组
- ◀ `scipy.stats.gaussian_kde()` 高斯核密度估计
- ◀ `statsmodels.api.nonparametric.KDEUnivariate()` 构造一元 KDE

19.1 似然概率：给定分类条件下的概率密度

本章也是采用鸢尾花数据对鸢尾花分类进行预测；不同的是，本章采用花萼长度、花萼宽度两个特征，相当于上一章贝叶斯分类的“升维”。本章和上一章的编排类似，请大家对照阅读；因此，这两章也共享一个知识导图。

为了估算 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$ ，首先提取标签为 C_1 (Setosa) 的 50 个样本，根据样本所在具体位置利用高斯 KDE 估计 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$ 。

图 1 所示为通过高斯 KDE 方法估算得到的似然概率 PDF 曲面 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$ 。 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$ 和水平面包裹的几何体的体积为 1。标签为 C_1 的鸢尾花数据，花萼长度主要集中在 4.5 ~ 5.5 cm 区域，花萼宽度则集中在 3 ~ 4 cm 区域。这个区域的 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$ 曲面高度最高，也就是可能性最大。

本书第 6 章还给出过条件概率 $f_{X_1, X_2|Y}(x_1, x_2 | y = C_1)$ 平面等高线和条件边缘概率密度曲线，请大家回顾。

⚠ 注意，要计算概率，需要对 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$ 进行二重积分。对 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$ “偏积分”的结果为条件边缘概率密度 $f_{X_1|Y}(x_1|C_1)$ 或 $f_{X_2|Y}(x_2|C_1)$ 。

图 2 所示为似然概率 $f_{X_1, X_2|Y}(x_1, x_2|C_2)$ 曲面。图 3 为似然概率 $f_{X_1, X_2|Y}(x_1, x_2|C_3)$ 曲面。

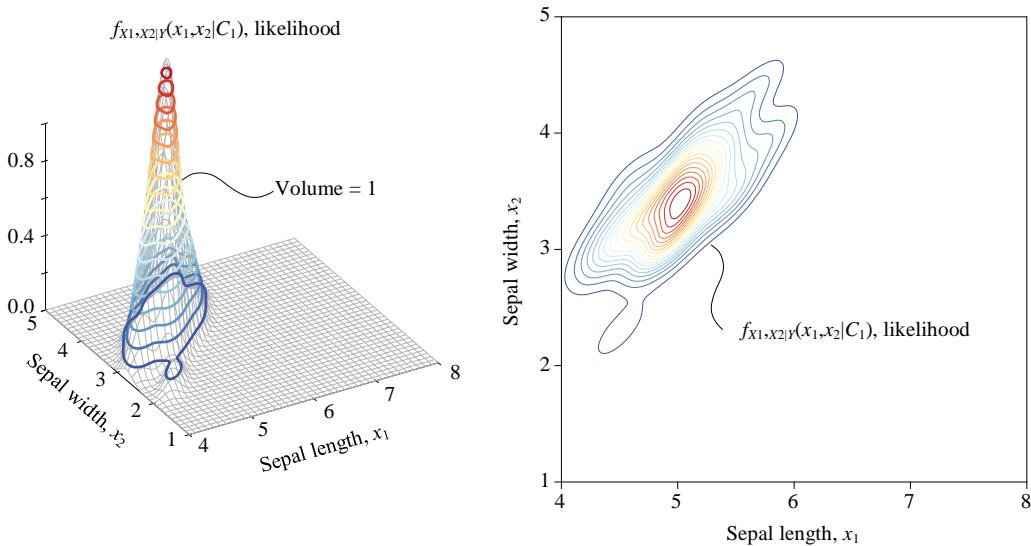
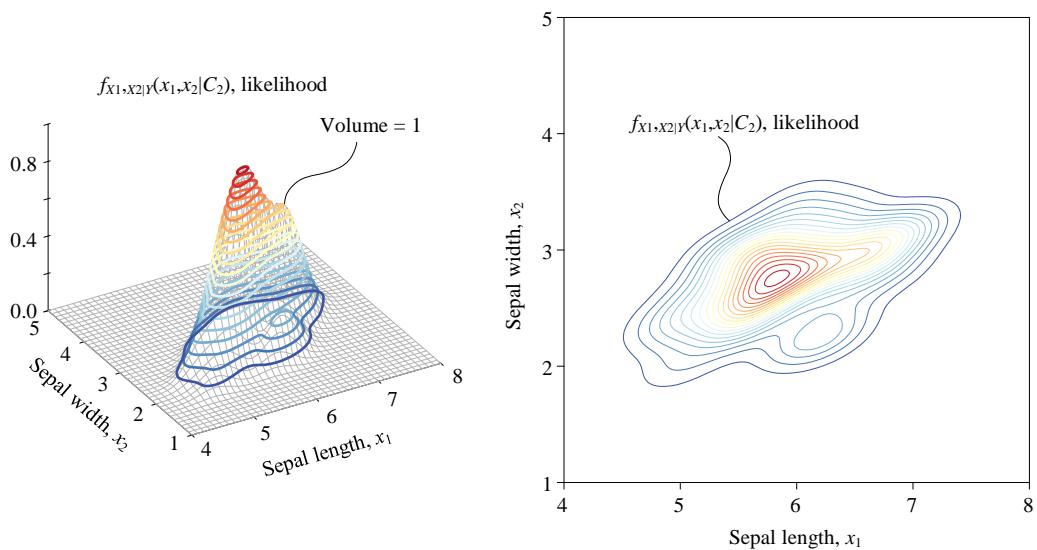
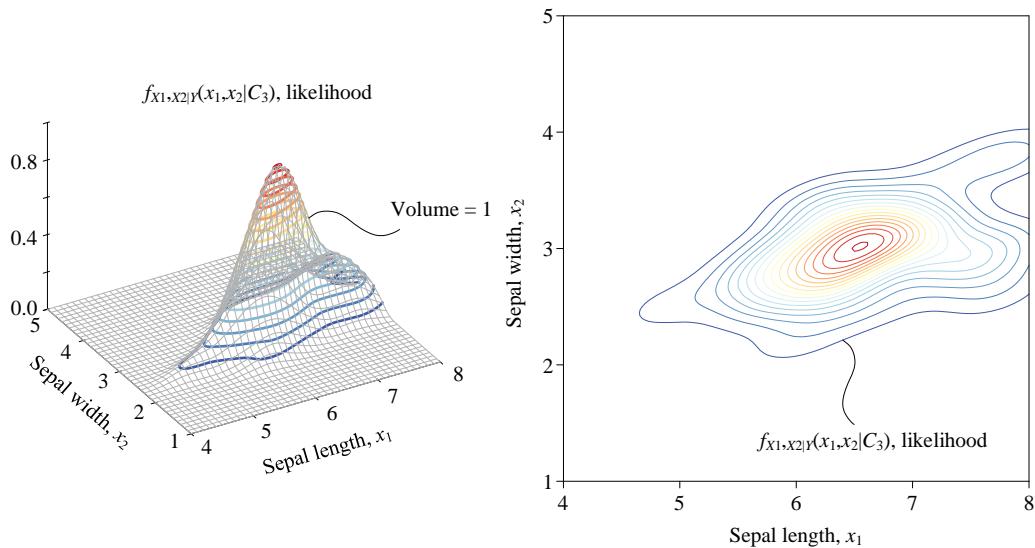


图 1. 似然概率 PDF 曲面 $f_{X_1, X_2|Y}(x_1, x_2|C_1)$

图 2. 似然概率 PDF 曲面 $f_{X1,X2|Y}(x_1,x_2|C_2)$ 图 3. 似然概率 PDF 曲面 $f_{X1,X2|Y}(x_1,x_2|C_3)$

比较

图 4 比较 $f_{X1,X2|Y}(x_1,x_2|C_1)$ 、 $f_{X1,X2|Y}(x_1,x_2|C_2)$ 、 $f_{X1,X2|Y}(x_1,x_2|C_3)$ 三个似然概率平面等高线。

本章计算先验概率的方式和上一章完全一致，请大家回顾。然后利用贝叶斯定理，根据似然概率和先验概率就可以计算联合概率和证据因子。

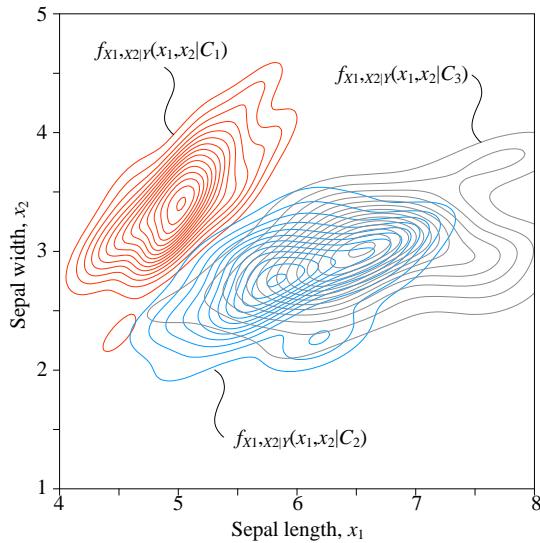


图 4. 比较三个似然概率曲面，平面等高线

19.2 联合概率：可以作为分类标准

联合概率 $f_{X1,X2,Y}(x_1, x_2, C_k)$ 描述三个事件 $X_1 = x_1$ 、 $X_2 = x_2$ 、 $Y = C_k$ 同时发生可能性。

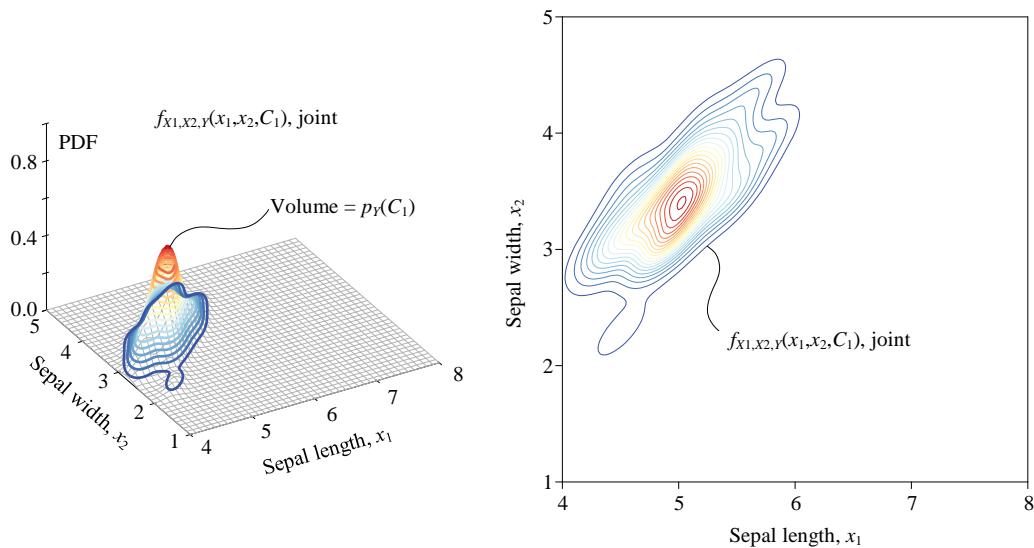
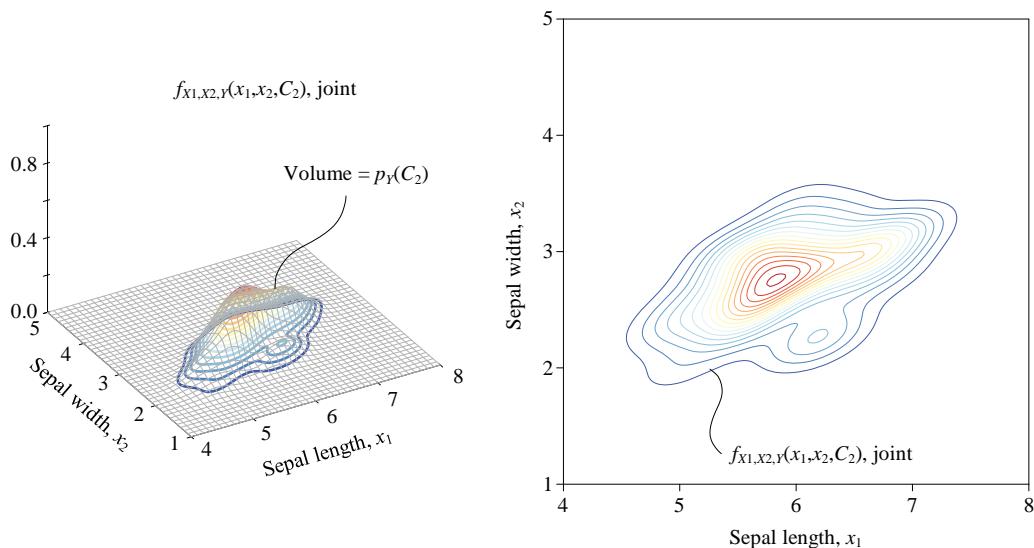
根据贝叶斯定理，联合概率 $f_{X1,X2,Y}(x_1, x_2, C_k)$ 可以通过似然概率 $f_{X1,X2|Y}(x_1, x_2 | C_k)$ 和先验概率 $p_Y(C_k)$ 相乘得到：

$$\overbrace{f_{X1,X2,Y}(x_1, x_2, C_k)}^{\text{Joint}} = \overbrace{f_{X1,X2|Y}(x_1, x_2 | C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}} \quad (1)$$

对于鸢尾花分类问题， Y 为离散随机变量，而先验概率 $p_Y(C_k)$ 本身为概率质量函数， $p_Y(C_k)$ 在(1)中仅仅起到缩放作用。

图 5 所示为联合概率 PDF 曲面 $f_{X1,X2,Y}(x_1, x_2, C_1)$ 、 $f_{X1,X2,Y}(x_1, x_2, C_2)$ 和水平面包裹的几何体的体积为 $p_Y(C_1)$ 。图 6 和图 7 所示为 $f_{X1,X2,Y}(x_1, x_2, C_2)$ 和 $f_{X1,X2,Y}(x_1, x_2, C_3)$ 两个联合概率曲面。

上一章介绍过，比较三个联合概率曲面高度可以用作鸢尾花分类预测的依据。

图 5. 联合概率 PDF 曲面 $f_{x_1,x_2,y}(x_1,x_2,C_1)$ 图 6. 联合概率 PDF 曲面 $f_{x_1,x_2,y}(x_1,x_2,C_2)$

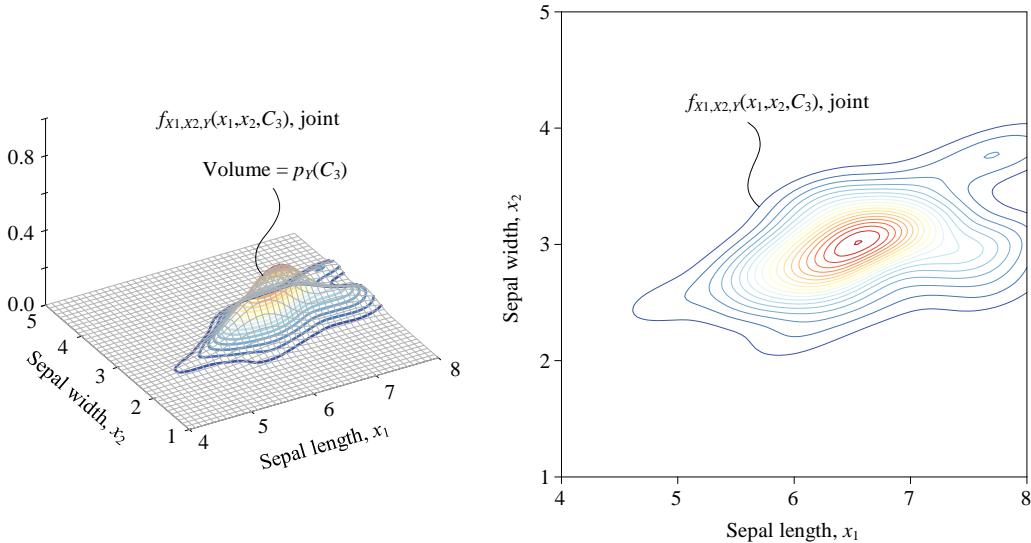
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 7. 联合概率 PDF 曲面 $f_{X1,X2,Y}(x_1, x_2, C_3)$

19.3 证据因子：和分类无关

证据因子 $f_{X1,X2}(x_1, x_2)$ 描述样本数据的分布情况，和分类无关。

C_1, C_2, C_3 为一组不相容分类，对鸢尾花数据样本空间 Ω 形成分割。根据全概率定理，下式成立：

$$\overbrace{f_{X1,X2}(x_1, x_2)}^{\text{Evidence}} = \sum_{k=1}^3 \overbrace{f_{X1,X2,Y}(x_1, x_2, C_k)}^{\text{Joint}} = \sum_{k=1}^3 \overbrace{f_{X1,X2|Y}(x_1, x_2 | C_k)}^{\text{Likelihood}} \overbrace{p_Y(C_k)}^{\text{Prior}} \quad (2)$$

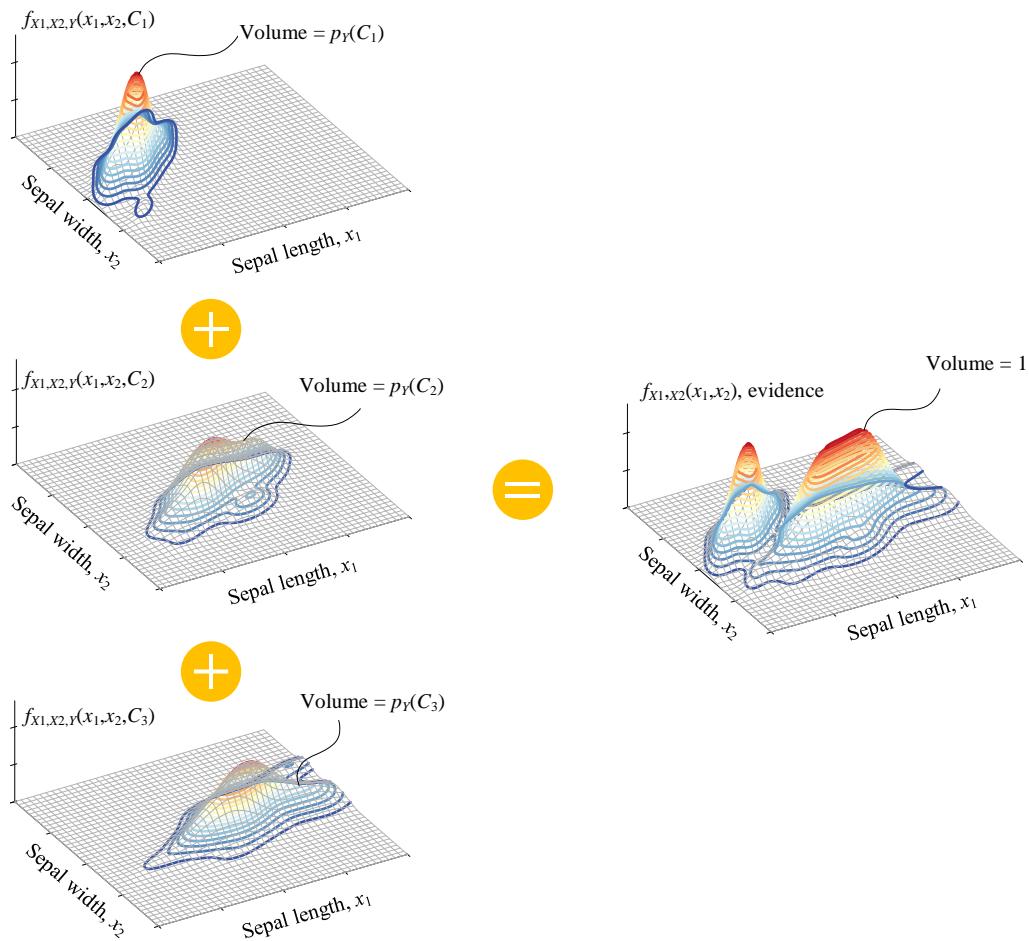
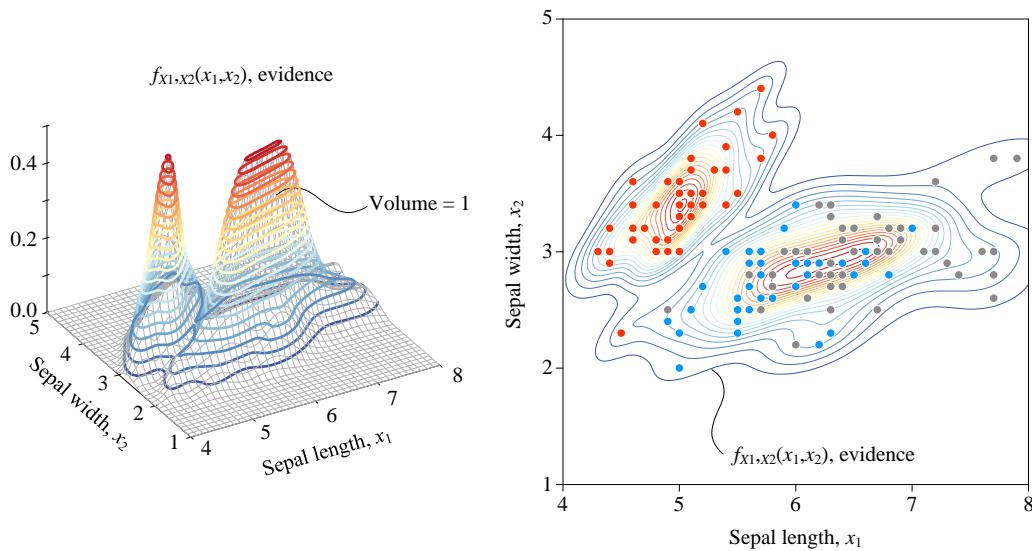
上式可以用来估算 $f_{X1,X2}(x_1, x_2)$ 。

把 (2) 展开来写，证据因子 $f_{X1,X2}(x_1, x_2)$ 可以通过下式计算得到：

$$\begin{aligned} f_{X1,X2}(x_1, x_2) &= f_{X1,X2,Y}(x_1, x_2, C_1) + f_{X1,X2,Y}(x_1, x_2, C_2) + f_{X1,X2,Y}(x_1, x_2, C_3) \\ &= f_{X1,X2|Y}(x_1, x_2 | C_1) p_Y(C_1) + f_{X1,X2|Y}(x_1, x_2 | C_2) p_Y(C_2) + f_{X1,X2|Y}(x_1, x_2 | C_3) p_Y(C_3) \end{aligned} \quad (3)$$

图 8 所示为叠加联合概率 PDF 曲面，计算证据因子 PDF 的过程。图 8 左侧三个几何体的体积分别为 $p_Y(C_1)$ 、 $p_Y(C_2)$ 、 $p_Y(C_3)$ 。显然 $p_Y(C_1)$ 、 $p_Y(C_2)$ 、 $p_Y(C_3)$ 三者之和为 1。

图 9 所示为 $f_{X1,X2}(x_1, x_2)$ 的曲面和平面等高线图。可以发现 $f_{X1,X2}(x_1, x_2)$ 较好地描述了样本数据分布。

图 8. 叠加联合概率曲面，估算证据因子概率密度函数 $f_{x1,x2}(x_1, x_2)$ 图 9. $f_{x1,x2}(x_1, x_2)$ 曲面及平面等高线

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

19.4 后验概率：也是分类的依据

$f_{Y|X_1,X_2}(C_k | x_1, x_2)$ 作为条件概率，指的是在 $X_1 = x_1$ 和 $X_2 = x_2$ 发生条件下，事件 $Y = C_k$ 发生的概率。上一章提到， $f_{Y|X_1,X_2}(C_k | x_1, x_2)$ 本身为概率，也就是说 $f_{Y|X_1,X_2}(C_k | x_1, x_2)$ 的取值范围为 $[0, 1]$ ；因此，后验概率 $f_{Y|X_1,X_2}(C_k | x_1, x_2)$ 又叫成员值。

根据贝叶斯定理，当 $f_{X_1,X_2}(x_1, x_2) > 0$ 时，后验概率 PDF $f_{Y|X_1,X_2}(C_k | x_1, x_2)$ 可以根据下式计算得到：

$$\overbrace{f_{Y|X_1,X_2}(C_k | x_1, x_2)}^{\text{Posterior}} = \frac{\overbrace{f_{X_1,X_2,Y}(x_1, x_2, C_k)}^{\text{Joint}}}{\overbrace{f_{X_1,X_2}(x_1, x_2)}^{\text{Evidence}}} \quad (4)$$

图 10、图 11、图 12 所示分别为后验概率 $f_{Y|X_1,X_2}(C_1 | x_1, x_2)$ 、 $f_{Y|X_1,X_2}(C_2 | x_1, x_2)$ 、 $f_{Y|X_1,X_2}(C_3 | x_1, x_2)$ 对应曲面和平面等高线。

上一章提到，后验概率（成员值）存在以下关系：

$$\sum_{k=1}^3 \underbrace{f_{Y|X_1,X_2}(C_k | x_1, x_2)}_{\text{Posterior}} = 1 \quad (5)$$

这意味着，图 10、图 11、图 12 三幅图曲面叠加在一起得到高度为 1 的“平台”。

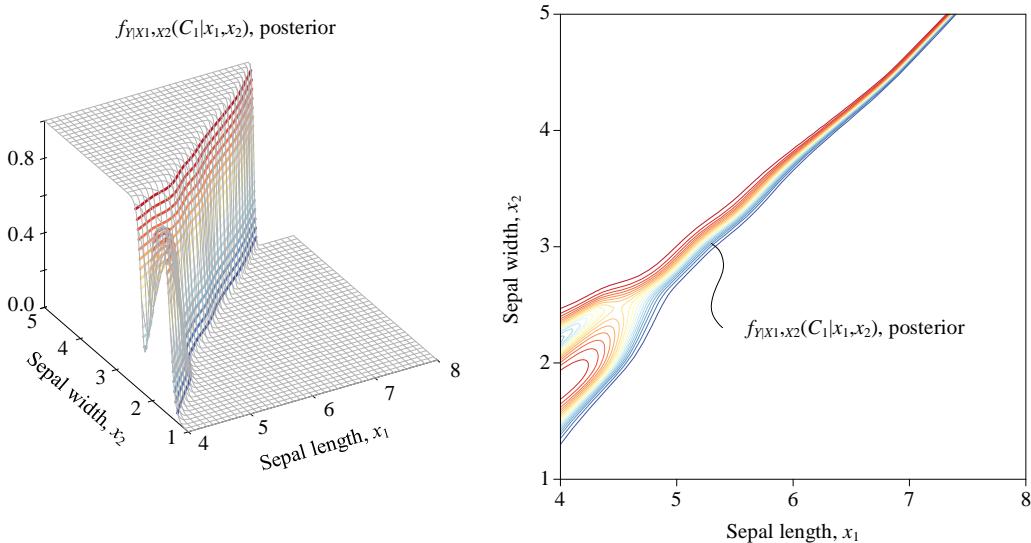
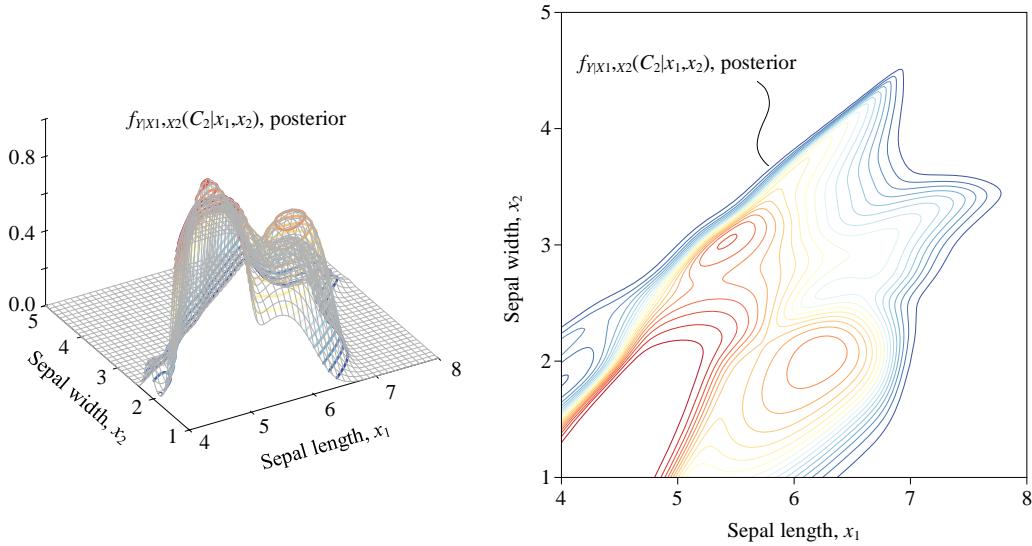
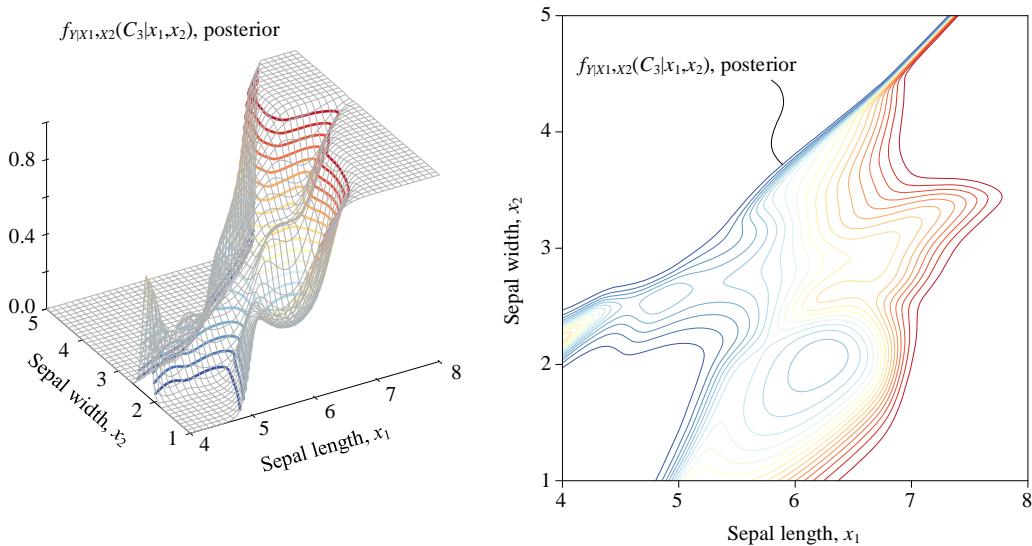


图 10. 后验概率 $f_{Y|X_1,X_2}(C_1 | x_1, x_2)$ 对应曲面和平面等高线

图 11. 后验概率 $f_{Y|X1,X2}(C_2 | x_1, x_2)$ 对应曲面和平面等高线图 12. 后验概率 $f_{Y|X1,X2}(C_3 | x_1, x_2)$ 对应曲面和平面等高线

分类依据

在给定任意花萼长度 x_1 和花萼宽度 x_2 的条件下，比较图 13 所示三个后验概率 $f_{Y|X1,X2}(C_1 | x_1, x_2)$ 、 $f_{Y|X1,X2}(C_2 | x_1, x_2)$ 、 $f_{Y|X1,X2}(C_3 | x_1, x_2)$ 大小，最大后验概率对应的标签就可以作为鸢尾花分类依据。

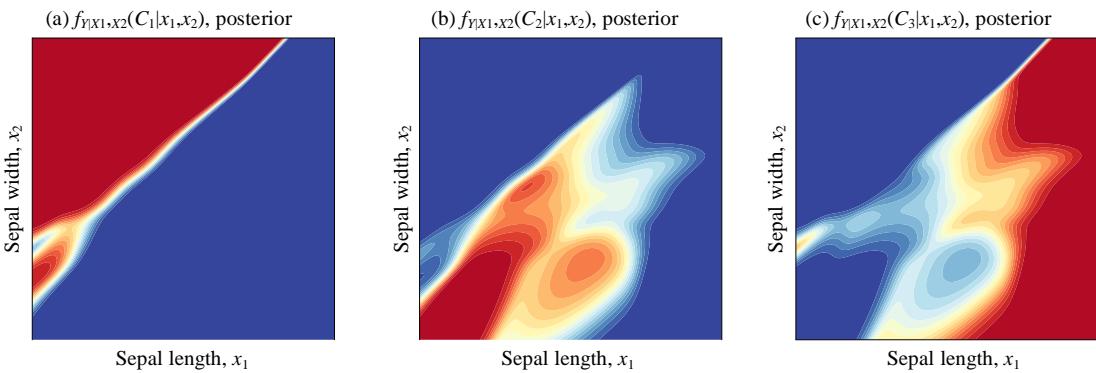


图 13. 比较三个后验概率曲面平面填充等高线

也就是说，这个分类问题对应的优化目标为最大化后验概率，即：

$$\hat{y} = \arg \max_{C_k} f_{Y|X1,X2}(C_k | x_1, x_2) \quad (6)$$

其中， $k = 1, 2, \dots, K$ 。对于鸢尾花三分类问题， $K = 3$ 。根据后验 \propto 似然 \times 先验，我们也可以最大化“似然 \times 先验”。

图 14 这幅图中曲线就是所谓**决策边界** (decision boundary)，决策边界将平面划分成三个区域，每个区域对应一类鸢尾花标签。

→ 《机器学习》一册将探讨更多分类算法。

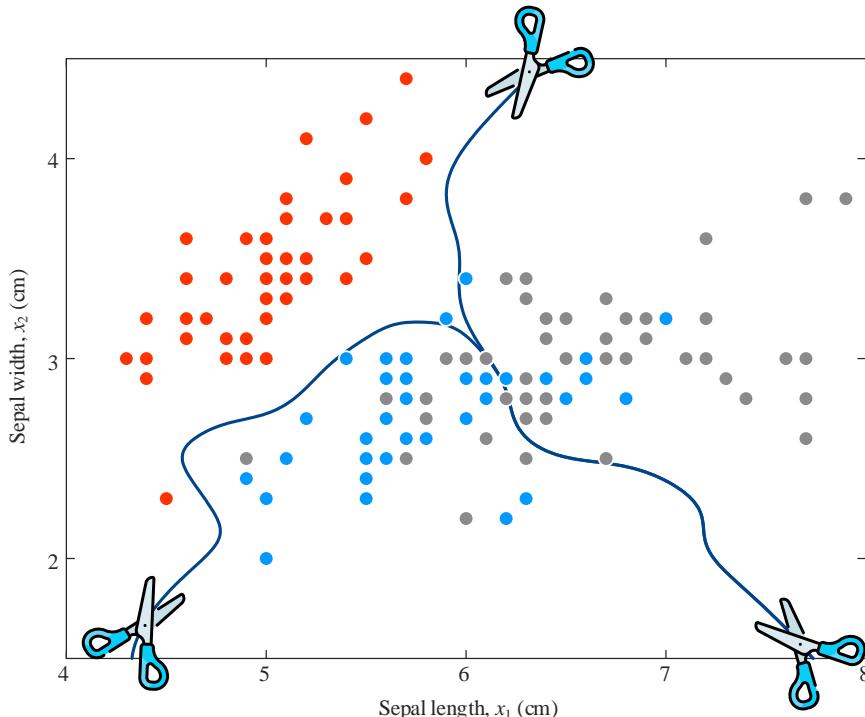


图 14. 朴素贝叶斯决策边界，基于核密度估计 KDE

19.5 独立：不代表条件独立

本章最后以鸢尾花数据为例再次区分“独立”和“条件独立”这两个概念。

如果假设鸢尾花萼长度 X_1 和萼宽 X_2 两个随机变量独立，联合概率 $f_{X_1, X_2}(x_1, x_2)$ 可以通过下式计算得到：

$$\underbrace{f_{X_1, X_2}(x_1, x_2)}_{\text{Joint}} = \underbrace{f_{X_1}(x_1)}_{\text{Marginal}} \cdot \underbrace{f_{X_2}(x_2)}_{\text{Marginal}} \quad (7)$$

图 15 所示为假设 X_1 和 X_2 独立时，估算得到的联合概率 $f_{X_1, X_2}(x_1, x_2)$ 曲面和平面等高线。观察图 15 等高线，容易发现假设 X_1 和 X_2 独立估算得到的联合概率 $f_{X_1, X_2}(x_1, x_2)$ 并没有很好地描述鸢尾花数分布。

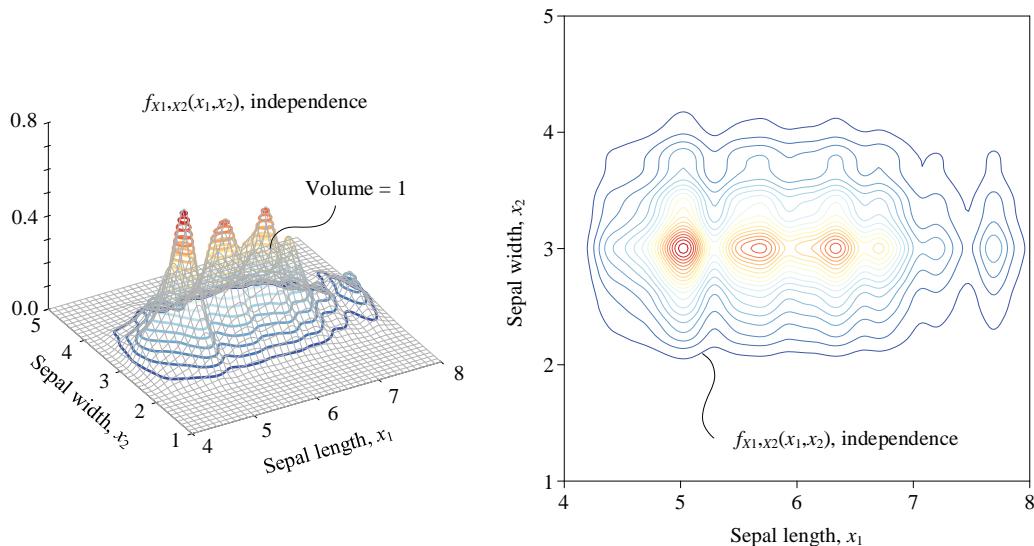
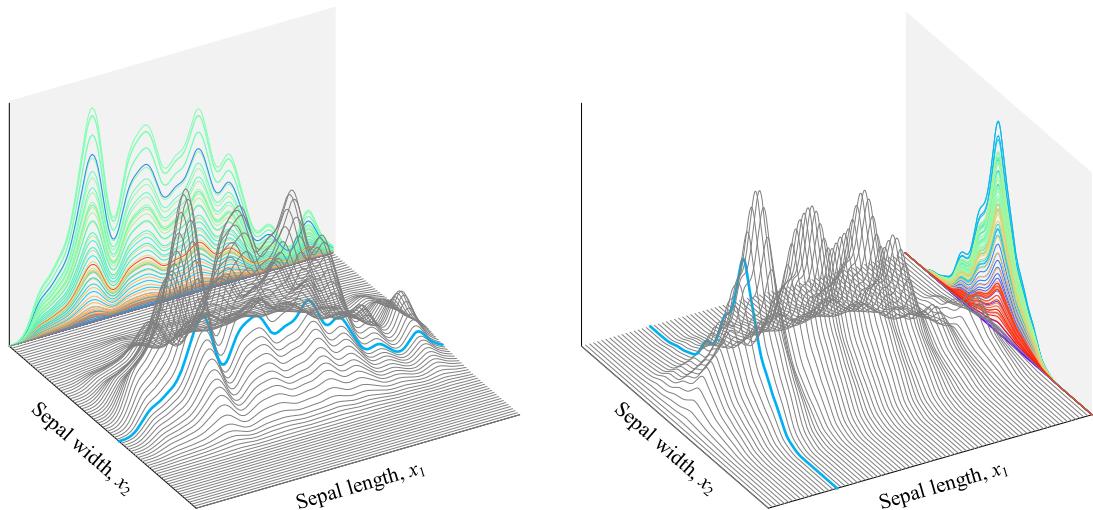
图 15. X_1 和 X_2 独立时，估算得到的联合概率 $f_{X_1, X_2}(x_1, x_2)$ 曲面和曲面等高线

图 16 所示为将 $f_{X_1, X_2}(x_1, x_2)$ 曲面在两个不同平面的投影。可以发现在不同平面上的投影都相当于该方向上边缘分布的高度上缩放。

图 16. $f_{X1,X2|Y}(x_1, x_2 | C_k)$ 曲面在两个不同平面的投影，假设特征独立

19.6 条件独立：不代表独立

回顾本书第 3 章讲过的条件独立。如果 $\Pr(A, B|C) = \Pr(A|C) \cdot \Pr(B|C)$ ，则称事件 A, B 对于给定事件 C 是条件独立的。也就是说，当 C 发生条件下， A 发生与否与 B 发生与否无关。

对于鸢尾花样本数据，给定 $Y = C_k$ 的条件下，如果假设花萼长度 X_1 、花萼宽度 X_2 条件独立，则下式成立：

$$f_{X1,X2|Y}(x_1, x_2 | C_k) = f_{X1|Y}(x_1 | C_k) \cdot f_{X2|Y}(x_2 | C_k) \quad (8)$$

上式相当于，一个类别、一个类别地分析数据。

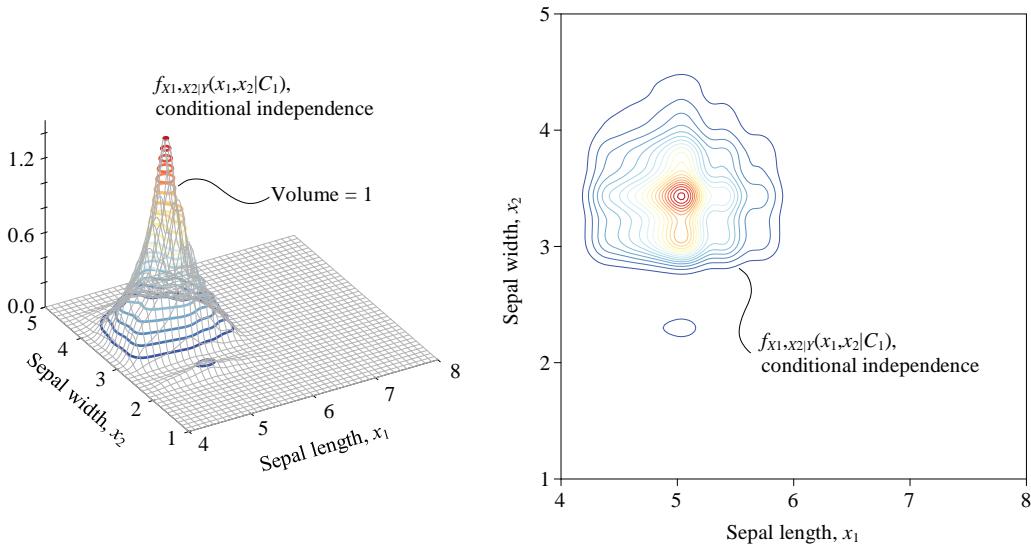
$Y = C_1$ 条件

给定 $Y = C_1$ 的条件下，如果假设 X_1, X_2 条件独立，则：

$$f_{X1,X2|Y}(x_1, x_2 | C_1) = f_{X1|Y}(x_1 | C_1) \cdot f_{X2|Y}(x_2 | C_1) \quad (9)$$

图 17 所示为在 $Y = C_1$ 的条件下，假设 X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1, x_2 | C_1)$ 。

本书第 6 章给出过假设条件独立情况下 $f_{X1,X2|Y}(x_1, x_2 | C_1)$ 、边缘似然概率 $f_{X1|Y}(x_1 | C_1)$ 、 $f_{X2|Y}(x_2 | C_1)$ 三者关系，请大家回顾。如果把 $f_{X1|Y}(x_1 | C_1)$ 、 $f_{X2|Y}(x_2 | C_1)$ 看做两个向量的话， $f_{X1,X2|Y}(x_1, x_2 | C_1)$ 就是两者的张量积。

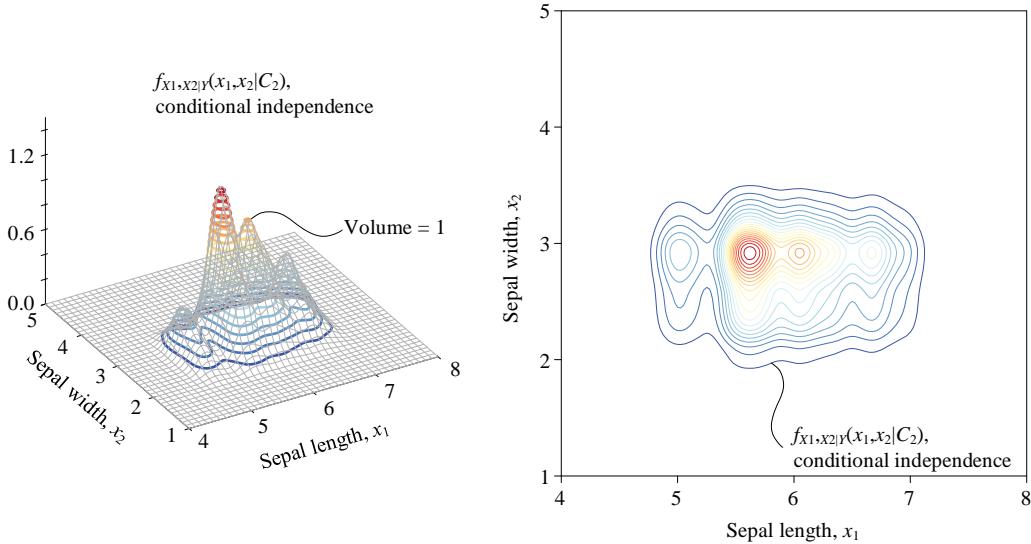
图 17. 在 $Y = C_1$ 的条件下， X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_1)$

$Y = C_2$ 条件

给定 $Y = C_2$ 的条件下，如果假设 X_1 、 X_2 条件独立，则：

$$f_{X1,X2|Y}(x_1,x_2|C_2) = f_{X1|Y}(x_1|C_2) \cdot f_{X2|Y}(x_2|C_2) \quad (10)$$

图 18 所示为在 $Y = C_2$ 的条件下，假设 X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_2)$ 。

图 18. 在 $Y = C_2$ 的条件下， X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X1,X2|Y}(x_1,x_2|C_2)$

$Y = C_3$ 条件

给定 $Y = C_3$ 的条件下，如果假设 X_1 、 X_2 条件独立，则：

$$f_{X_1, X_2|Y}(x_1, x_2|C_3) = f_{X_1|Y}(x_1|C_3) \cdot f_{X_2|Y}(x_2|C_3) \quad (11)$$

图 19 所示为在 $Y = C_3$ 的条件下，假设 X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X_1, X_2|Y}(x_1, x_2|C_3)$ 。

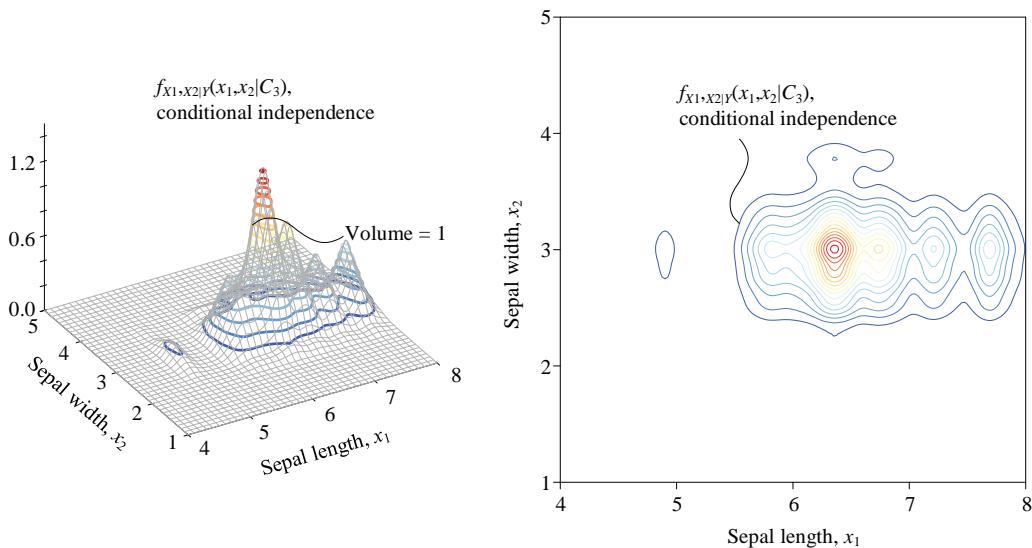


图 19. 在 $Y = C_3$ 的条件下， X_1 和 X_2 条件独立，估算得到的似然概率 $f_{X_1, X_2|Y}(x_1, x_2|C_3)$

估算证据因子

假设条件独立，证据因子 $f_{X_1, X_2}(x_1, x_2)$ 可以通过下式计算得到：

$$\begin{aligned} f_{X_1, X_2}(x_1, x_2) &= f_{X_1, X_2|Y}(x_1, x_2|C_1) \cdot p_Y(C_1) + f_{X_1, X_2|Y}(x_1, x_2|C_2) \cdot p_Y(C_2) + f_{X_1, X_2|Y}(x_1, x_2|C_3) \cdot p_Y(C_3) \\ &= f_{X_1|Y}(x_1|C_1) \cdot f_{X_2|Y}(x_2|C_1) \cdot p_Y(C_1) + \\ &\quad f_{X_1|Y}(x_1|C_2) \cdot f_{X_2|Y}(x_2|C_2) \cdot p_Y(C_2) + \\ &\quad f_{X_1|Y}(x_1|C_3) \cdot f_{X_2|Y}(x_2|C_3) \cdot p_Y(C_3) \end{aligned} \quad (12)$$

上式代表一种多元概率密度估算方法。图 20 所示为假设条件独立，估算 $f_{X_1, X_2}(x_1, x_2)$ 概率密度的过程。图 21 所示为 $f_{X_1, X_2}(x_1, x_2)$ 曲面和平面等高线。



条件独立这一假设对于朴素贝叶斯方法至关重要。《机器学习》一册将分别介绍朴素贝叶斯分类，和高斯朴素贝叶斯分类。

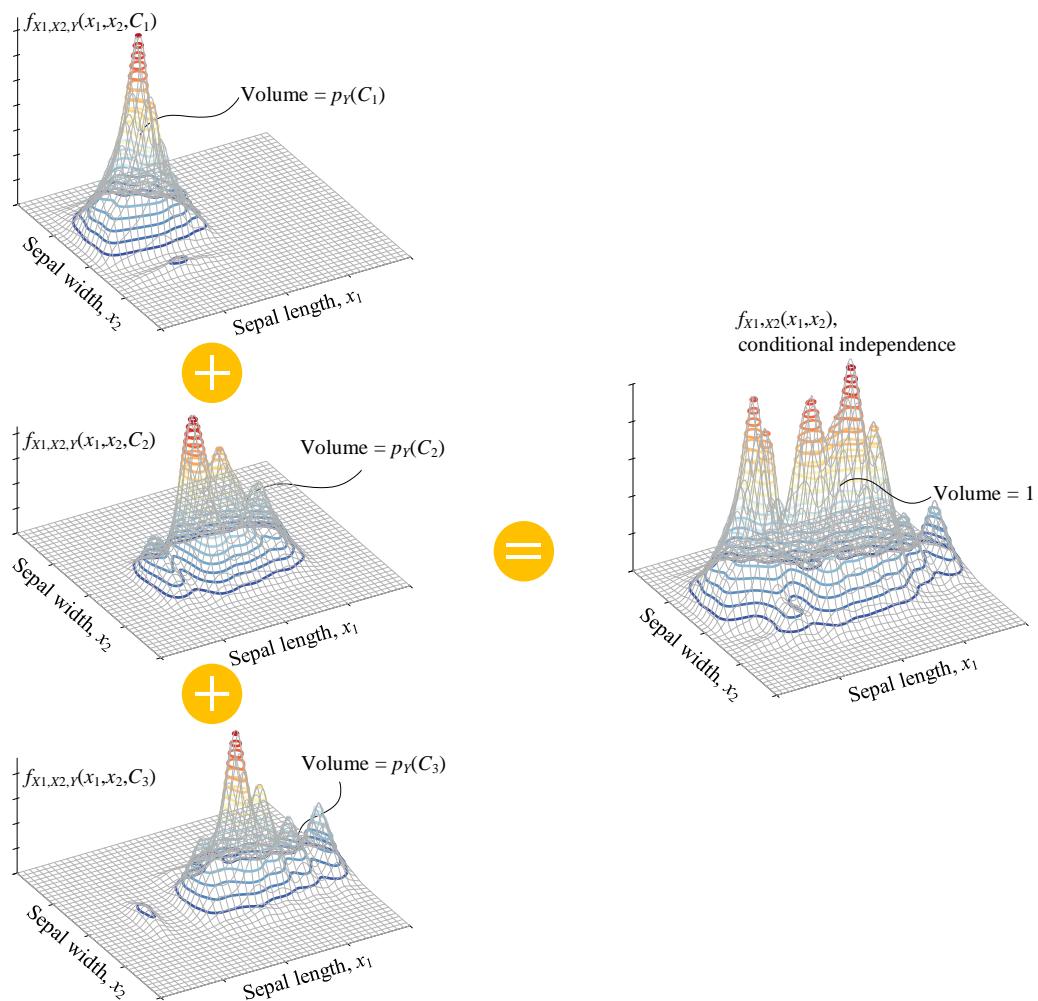
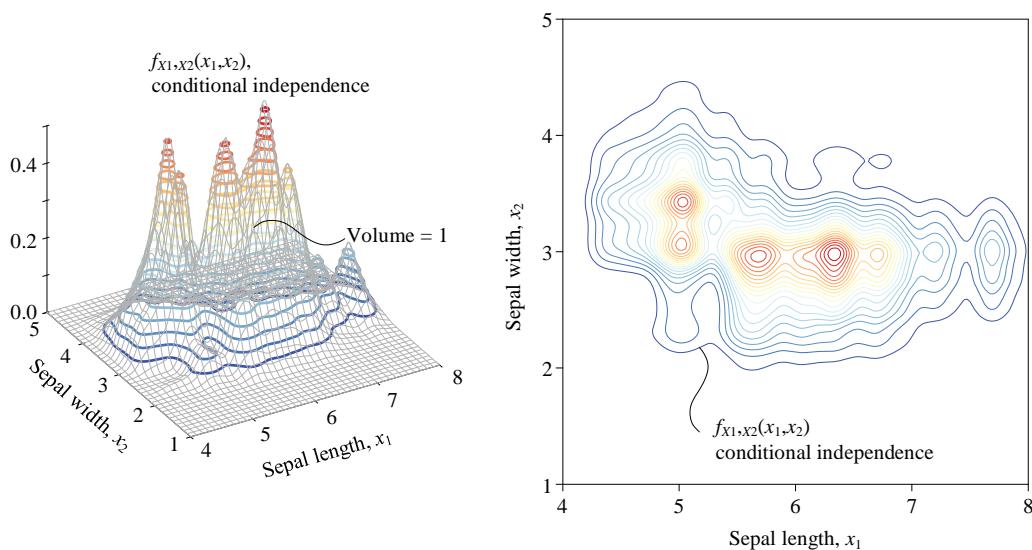
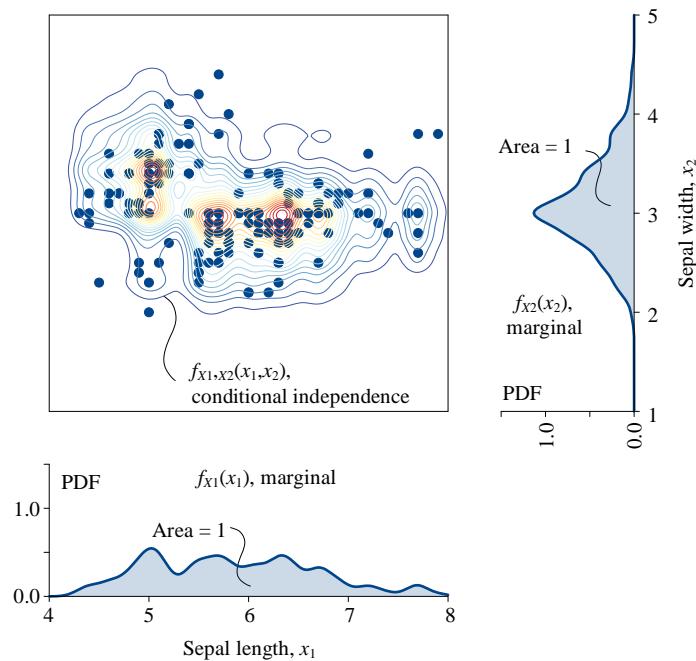
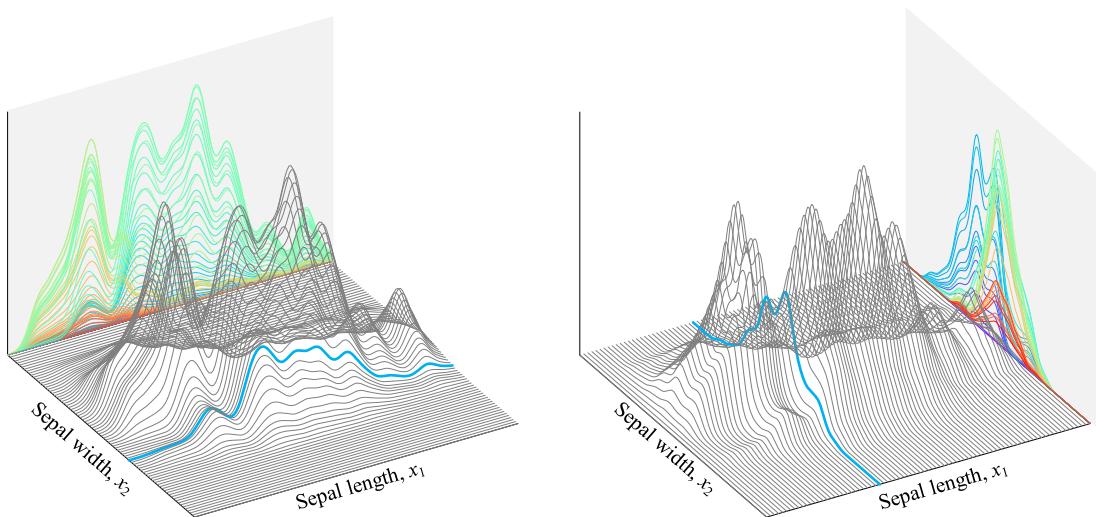
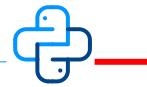
图 20. 假设条件独立，合成叠加得到证据因子 $f_{x_1, x_2}(x_1, x_2)$ 

图 21. 假设条件独立，证据因子 $f_{X_1, X_2}(x_1, x_2)$ 曲面和平面等高线

如图 22 所示，显然采用条件独立假设估算得到的证据因子概率密度函数 $f_{X_1, X_2}(x_1, x_2)$ 对样本数据分布的贴合度更高。图 23 所示为 $f_{X_1, X_2}(x_1, x_2)$ 在两个竖直平面上的投影，请大家对比图 16 分析。

图 22. 假设条件独立，证据因子 $f_{X_1, X_2}(x_1, x_2)$ 等高线，和边缘概率密度 $f_{X_1}(x_1)$ 、 $f_{X_2}(x_2)$ 曲线关系图 23. 假设条件独立，证据因子 $f_{X_1, X_2}(x_1, x_2)$ 曲面在两个平面投影曲线



Bk5_Ch19_01.py 代码绘制本章绝大部分图像。



贝叶斯分类是一种基于贝叶斯定理的分类方法，它根据给定的特征和类别之间的关系，通过学习训练数据集中的先验概率和条件概率，对新的输入进行分类。贝叶斯分类将输入数据看作特征向量，并根据这些特征向量的先验概率和条件概率来计算其属于不同类别的后验概率，最终选择概率最大的类别作为输出。

贝叶斯分类的优点在于其能够处理高维数据，并且在数据量较小的情况下表现良好，同时还能处理具有噪声或缺失数据的情况。《机器学习》一册将专门介绍贝叶斯分类的特殊形式——朴素贝叶斯分类。

下面三章，我们把贝叶斯定理应用到贝叶斯推断中。

20

Bayesian Inference 101

贝叶斯推断入门

参数不确定，参数对应概率分布



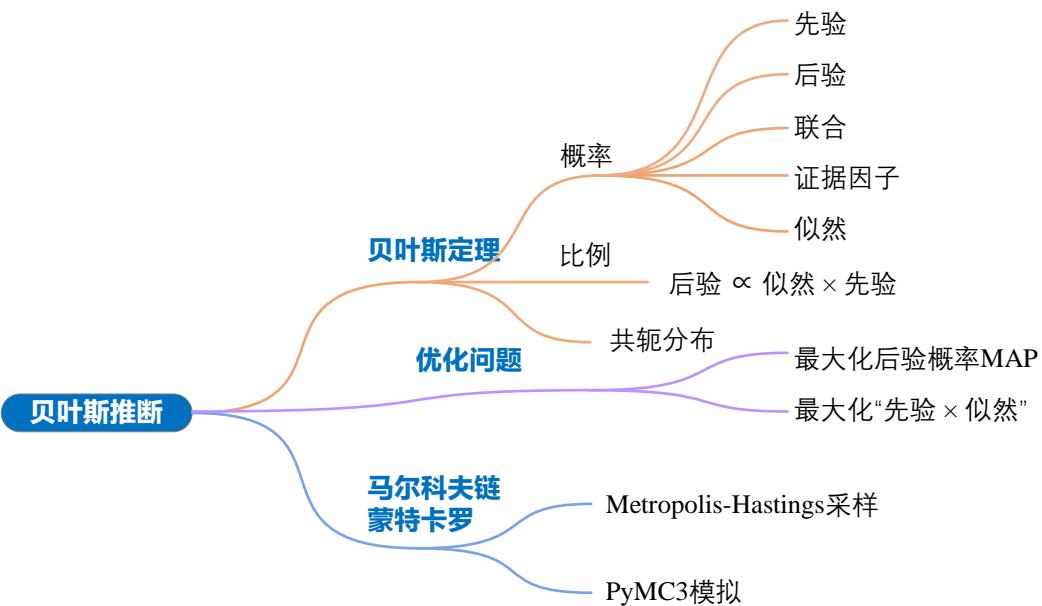
没有事实，只有解释。

There are no facts, only interpretations.

—— 弗里德里希·尼采 (Friedrich Nietzsche) | 德国哲学家 | 1844 ~ 1900



- ◀ matplotlib.pyplot.axvline() 绘制竖直线
- ◀ matplotlib.pyplot.fill_between() 区域填充颜色
- ◀ numpy.cumsum() 累加
- ◀ scipy.stats.bernoulli.rvs() 满足伯努利分布的随机数
- ◀ scipy.stats.beta() Beta 分布



20.1 贝叶斯推断：更贴合人脑思维

一个让人“头大”的公式

本章和下一章的关键就是如何理解、应用以下公式进行贝叶斯推断：

$$f_{\Theta|x}(\theta|x) = \frac{f_{x|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int f_{x|\Theta}(x|\vartheta)f_{\Theta}(\vartheta)d\vartheta} \quad (1)$$

值得注意的是这个公式还有如下常见的几种其他写法：

$$\begin{aligned} f_{\Theta|x}(\theta|x) &= \frac{f_{x|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int_{\theta'} f_{x|\Theta}(x|\theta')f_{\Theta}(\theta')d\theta'} \\ f_{\Theta|x}(\theta|x) &= \frac{f_{x|\Theta}(x|\theta)g_{\Theta}(\theta)}{\int_{\vartheta} f_{x|\Theta}(x|\vartheta)g_{\Theta}(\vartheta)d\vartheta} \\ p_{\Theta|x}(\theta|x) &= \frac{p_{x|\Theta}(x|\theta)p_{\Theta}(\theta)}{\int_{\theta'} p_{x|\Theta}(x|\theta')p_{\Theta}(\theta')d\theta'} \end{aligned} \quad (2)$$

有些书有把 x 写成 y 情况，也有用 $\pi()$ 代表概率密度/质量分布函数。总而言之，(1) 的表达方式很多，大家见多了，也就“见怪不怪”了。

(1) 这个公式是横在大家理解掌握贝叶斯推断之路上的一块“巨石”。本章试图用最简单的例子帮大家敲碎这块“巨石”。

在正式介绍这个公式之前，本节先用白话聊聊什么是**贝叶斯推断** (Bayesian inference)。

贝叶斯推断

本书第 16 章介绍过，在贝叶斯学派眼里，模型参数本身也是随机变量，也服从某种分布。贝叶斯推断的核心就是，在以往的经验（先验概率）基础上，结合新的数据，得到新的概率（后验概率）。而模型参数分布随着外部样本数据不断输入而迭代更新。不同的是，频率派只考虑样本数据本身，不考虑先验概率。

依我看来，人脑的运作方式更贴近贝叶斯推断。

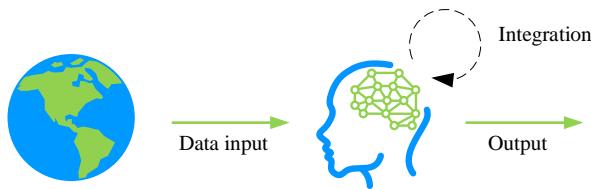


图 1. 人脑更像是一个贝叶斯推断机器

举个最简单的例子，试想你一早刚出门的时候发现忘带手机，大脑第一反应是——手机最可能在哪？

这个“贝叶斯推断”的结果一般基于两方面因素：一方面，日复一日的“找手机”的经验；另一方面，“今早、昨晚在哪用过手机”的最新数据。



图 2. 找手机

而且在不断寻找手机的过程，大脑不断提出“下一个最有可能的地点”。

比如，昨晚睡觉前刷了一小时手机，手机肯定在床上！

跑到床头，发现手机不在床上，那很可能在马桶附近，因为早晨方便的时候一般也会刷手机！

竟然也不在马桶附近！那最可能在沙发茶几上，因为坐着看电视的时候我也爱刷手机 …

试想，如果大脑没有以上“经验 + 最新数据”，你会怎么找手机？或者，“贝叶斯推断”找手机无果的时候，我们又会怎么办？

我们很可能会像“扫地机器人”一样，“逐点扫描”，把整个屋子从里到外歇斯底里地翻一遍。这种地毯式“采样”就类似频率派的做法。

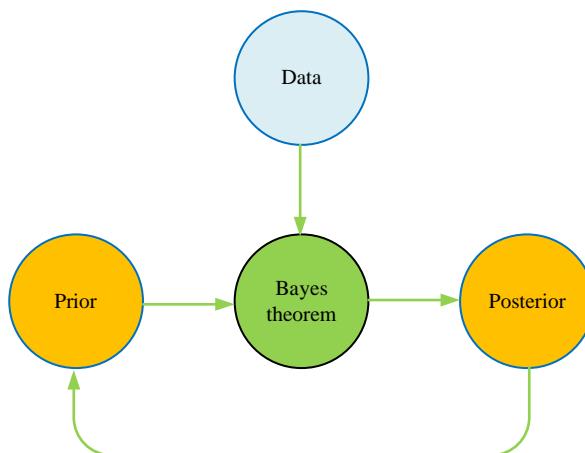


图 3. 通过贝叶斯定理迭代学习

这个找手机的过程也告诉我们，贝叶斯推断常常迭代使用。在引入新的样本数据后，先验概率产生后验概率。而这个后验概率也可以作为新的先验概率，再根据最新出现的数据，更新后验概率，如此往复。

人生来就是一个“学习机器”，“前事不忘后事之师”说的也是这个道理。通过不断学习（数据输入），我们不断更新自己对世界的认知（更新模型参数）。这个过程从出生一直持续到离开这个世界为止。

往大了说，人类认识世界的机制又何尝不是贝叶斯推断。在新的数据影响下，人类一次次创造、推翻、重构知识体系。这个过程循环往复，不断推动人类认知进步。

举个例子，统治西方世界思想界近千年的地心说被推翻后，日心说渐渐成了主流。在伽利略等一众巨匠的臂膀上，牛顿力学体系横空出世。在后世科学家不断努力完善下，牛顿力学体系和麦克斯韦电磁场理论为基础的物理大厦大功告成。当人们满心欢喜，以为物理学就剩下一些敲敲打打的修饰工作，结果蓝天之上又飘来了两朵乌云 …

20.2 从一元贝叶斯公式说起

先验

在任何引入任何观测数据之前，未知参数 θ 本身是随机变量，自身对应概率分布为 $f_\theta(\theta)$ ，这个分布叫做**先验分布** (prior distribution)。先验分布函数 $f_\theta(\theta)$ 中， θ 为随机变量， θ 是一个变量。 $\theta = \theta$ 代表随机变量 θ 的取值为 θ 。

似然

在 $\theta = \theta$ 条件下，观察到的数据 X 的分布为**似然分布** (likelihood distribution) $f_{X|\theta}(x|\theta)$ 。似然分布是一个条件概率。当 $\theta = \theta$ 取不同值时，似然分布 $f_{X|\theta}(x|\theta)$ 也有相应变化。

→ 回顾本书第 17 章介绍最大似然估计 MLE，优化问题的目标函数本质上就是似然函数 $f_{X|\theta}(x|\theta)$ 的连乘。第 17 章不涉及贝叶斯推断，因此我们没有用条件概率 $f_{X|\theta}(x|\theta)$ ，用的是 $f_X(x; \theta)$ 。**对数似然** (log-likelihood function) 就是对似然函数取对数，将连乘变成连加。

联合

根据贝叶斯定理， X 和 θ 的**联合分布** (joint distribution) 为：

$$\underbrace{f_{X,\Theta}(x, \theta)}_{\text{Joint}} = \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} \quad (3)$$

⚠ 请大家注意，为了方便，在贝叶斯推断中，我们不再区分概率密度函数 PDF、概率质量函数 PMF，所有概率分布均用 $f()$ 记号。而且，(1) 的分母也仅仅用积分符号。

证据

如果 X 为连续随机变量， X 的边缘概率分布为：

$$\underbrace{f_X(x)}_{\text{Evidence}} = \int_{\theta} \underbrace{f_{X,\Theta}(x, \theta)}_{\text{Joint}} d\theta = \int_{\theta} \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} d\theta \quad (4)$$

联合分布 $f_{X|\theta}(x|\theta)$ 对 θ “偏积分”消去了 θ ，积分结果 $f_X(x)$ 和 θ 无关。我们一般也管 $f_X(x)$ 叫做**证据因子** (evidence)，这和前两章的叫法一致。

$f_X(x)$ 和 θ 无关，这意味着观测到的数据对先验的选择没有影响。

后验

给定 $X=x$ 条件下， θ 的条件概率为：

$$f_{\Theta|X}(\theta|x) = \frac{\overbrace{f_{X,\Theta}(x, \theta)}^{\text{Joint}}}{\underbrace{f_X(x)}_{\text{Evidence}}} = \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{\int_{\theta} \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} d\theta} \quad (5)$$

⚠ 为了避免混淆，上式分母中用了花写 θ 。

$f_{\Theta|X}(\theta|x)$ 叫**后验分布** (posterior distribution)，它代表在整合“先验 + 样本数据”之后，我们对参数 θ 的新的“认识”。在连续迭代贝叶斯学习中，这个后验概率分布是下一个迭代的先验概率分布。

正比关系

通过前两章的学习，我们知道后验与先验和似然乘积成正比：

$$\underbrace{f_{\Theta|x}(\theta|x)}_{\text{Posterior}} \propto \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_{\Theta}(\theta)}_{\text{Prior}} \quad (6)$$

即，后验 \propto 似然 \times 先验。

但是为了得出真正的后验概率密度，本章的例子中我们还是要完成 $\int_{\theta} f_{X|\Theta}(x|\theta) f_{\Theta}(\theta) d\theta$ 积分。

 此外，这个积分很可能没有解析解（闭式解），可能需要用到数值积分或蒙特卡洛模拟。这是本书第 22 章要讲解的内容之一。

⚠ 注意，先验分布、后验分布是关于模型参数的分布。此外，通过一定的转化，我们可以把似然函数也变成有关模型参数的“分布”。

下面，我们便结合实例讲解贝叶斯推断。

20.3 走地鸡兔：比例完全不确定

回到本书第 16 章“鸡兔同笼”的例子。一个巨大无比农场散养大量“走地”鸡、兔。但是，农夫自己也说不清楚鸡兔的比例。

用 θ 代表兔子的比例随机变量，这意味着 θ 的取值范围为 $[0, 1]$ 。即， $\theta = 0.5$ 意味着农场有 50% 兔、50% 鸡， $\theta = 0.3$ 意味着有 30% 兔、70% 鸡。

为了搞清楚农场鸡兔比例，农夫决定随机抓 n 只动物。 X_1, X_2, \dots, X_n 为每次抓取动物的结果。 $X_i (i = 1, 2, \dots, n)$ 的样本空间为 $\{0, 1\}$ ，其中 0 代表鸡，1 代表兔。

⚠ 注意，抓取动物过程，我们同样忽略这对农场整体动物总体比例的影响。

先验

由于农夫完全不确定鸡兔的比例，我们选择连续均匀分布 $\text{Uniform}(0, 1)$ 为先验分布，所以 $f_{\Theta}(\theta)$ 为：

$$f_{\Theta}(\theta) = 1, \quad \theta \in [0, 1] \quad (7)$$

再次强调，先验分布代表我们对模型参数的“主观经验”，先验分布的选择独立于“客观”样本数据。

图 4 所示为 $[0, 1]$ 区间上的均匀分布，也就是说兔子比例 θ 可以是 $[0, 1]$ 区间内的任意一个数，而且可能性相同。

这个例子告诉我们，没有先验信息，或者先验分布不清楚，也不要紧！我们可以用常数或均匀分布作为先验分布。这种情况也叫**无信息先验** (uninformative prior)。

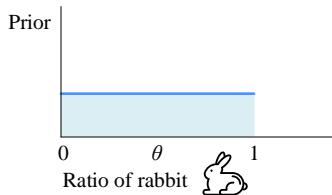


图 4. 选择连续均匀分布作为先验分布

似然

给定 $\theta = \theta$ 条件下， $X_1, X_2 \dots X_n$ 服从 IID 的伯努利分布 $Bernoulli(\theta)$ ，即：

$$\underbrace{f_{X_i|\Theta}(x_i | \theta)}_{\text{Likelihood}} = \theta^{x_i} (1-\theta)^{1-x_i} \quad (8)$$

其中， $\theta = \theta$ 代表农场中兔子的比例，取值范围为 $[0, 1]$ 区间任意数值； $1 - \theta$ 代表鸡的比例。 $X_i = x_i$ 代表某一次抓到的动物，0 代表鸡，1 代表兔。

也就是说，(8) 中， θ 是未知量。实际上，上式中似然概率 $f_{X_i|\Theta}(x_i | \theta)$ 代表概率质量函数。

本书前文提过，IID 的含义是**独立同分布** (Independent Identically Distribution)。在随机过程中，任何时刻的取值都为随机变量，如果这些随机变量服从同一分布，并且互相独立，那么这些随机变量是独立同分布。

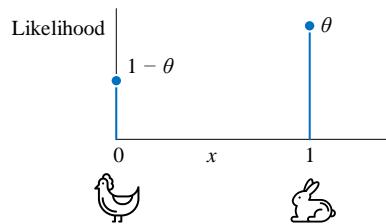


图 5. 似然分布

联合

因此， $X_1, X_2 \dots X_n, \theta$ 联合分布为：

$$\begin{aligned}
 f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta) &= \underbrace{f_{X_1, X_2, \dots, X_n | \Theta}(x_1, x_2, \dots, x_n | \theta)}_{\text{Likelihood}} \underbrace{f_{\Theta}(\theta)}_{\text{Prior}} \\
 &= f_{X_1 | \Theta}(x_1 | \theta) \cdot f_{X_2 | \Theta}(x_2 | \theta) \cdots f_{X_n | \Theta}(x_n | \theta) \cdot \underbrace{f_{\Theta}(\theta)}_1 \\
 &= \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}
 \end{aligned} \tag{9}$$

令：

$$s = \sum_{i=1}^n x_i \tag{10}$$

s 的含义是 n 次抽取中兔子的总数。

这样 (9) 可以写成：

$$f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta) = \theta^s (1-\theta)^{n-s} \tag{11}$$

上式中， $n - s$ 代表 n 次抽取中鸡的总数。

证据

证据因子 $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ ，即 $f_X(x)$ ，可以通过 $f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta)$ 对 θ “偏积分”得到：

$$\begin{aligned}
 f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= \int_{\theta} f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta) d\theta \\
 &= \int_{\theta} \theta^s (1-\theta)^{n-s} d\theta
 \end{aligned} \tag{12}$$

以上积分相当于在 θ 维度上压缩，结果 $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ 和 θ 无关。

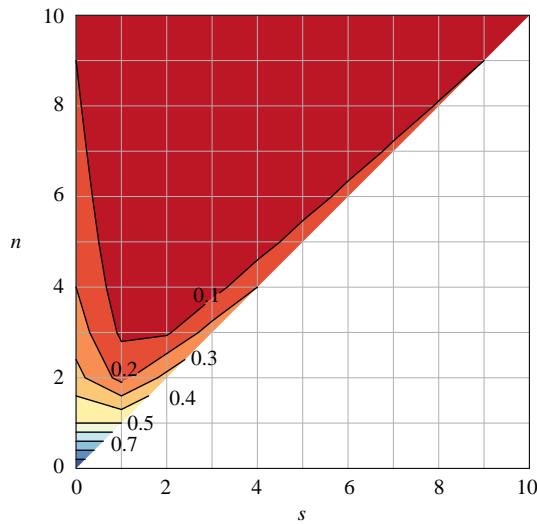
⚠ 再次强调，在贝叶斯推断中，上述积分很可能没有解析解。

想到本书第 7 章介绍的 Beta 函数，(12) 可以写成：

$$\begin{aligned}
 f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= \int_{\theta} \theta^{s+1-1} (1-\theta)^{n-s+1-1} d\theta \\
 &= B(s+1, n-s+1) = \frac{s!(n-s)!}{(n+1)!}
 \end{aligned} \tag{13}$$

利用 Beta 函数的性质，我们“逃过”积分运算。

图 6 所示为 $B(s+1, n-s+1)$ 函数随着 s 、 n 变化的平面等高线。

图 6. $B(s + 1, n - s + 1)$ 函数图像平面等高线

后验

由此，在 $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ 条件下， θ 的后验分布为：

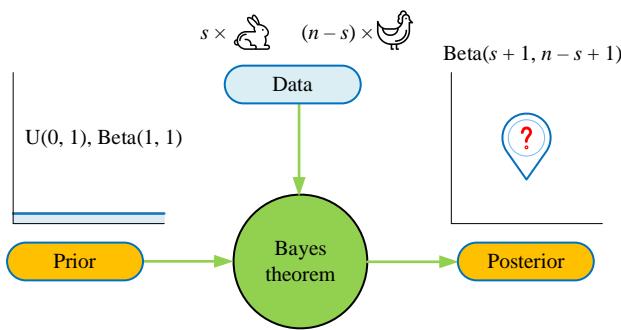
$$\begin{aligned}
 f_{\Theta|X_1, X_2, \dots, X_n}(\theta | x_1, x_2, \dots, x_n) &= \frac{\overbrace{f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta)}^{\text{Joint}}}{\underbrace{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)}_{\text{Evidence}}} \\
 &= \frac{\theta^s (1-\theta)^{n-s}}{B(s+1, n-s+1)} = \frac{\theta^{(s+1)-1} (1-\theta)^{(n-s+1)-1}}{B(s+1, n-s+1)}
 \end{aligned} \tag{14}$$

我们惊奇地发现，上式对应 $Beta(s + 1, n - s + 1)$ 分布。

总结来说，农夫完全不清楚鸡兔的比例，因此选择先验概率为 $Uniform(0, 1)$ 。抓取 n 只动物，知道其中有 s 只兔子， $n - s$ 只鸡，利用贝叶斯定理整合“先验概率 + 样本数据”得到后验概率为 $Beta(s + 1, n - s + 1)$ 分布。

⚠ 注意，实际上 $Uniform(0, 1)$ 就是 $Beta(1, 1)$ 。

马上，我们把蒙特卡罗模拟结果代入后验概率 $Beta(s + 1, n - s + 1)$ ，这样就可以看到后验分布的形状。

图 7. 先验 $U(0, 1)$ + 样本 $(s, n-s)$ → 后验 $Beta(s+1, n-s+1)$

正比关系

(14) 中分母 $B(s+1, n-s+1)$ 的作用是条件概率归一化。实际上，根据 (6)，我们只需要知道：

$$f_{\Theta|X_1, X_2, \dots, X_n}(\theta|x_1, x_2, \dots, x_n) \propto f_{X_1, X_2, \dots, X_n|\Theta}(x_1, x_2, \dots, x_n|\theta) f_\Theta(\theta) = \theta^s (1-\theta)^{n-s} \quad (15)$$

我们在前两章也看到了这个正比关系的应用。但是为了方便蒙特卡罗模拟，本节还是会使用 (14) 给出的后验分布解析式。

蒙特卡罗模拟

下面，我们编写 Python 代码来进行上述贝叶斯推断的蒙特卡洛模拟。先验分布为 $Uniform(0, 1)$ ，这意味着各种鸡兔比例可能性相同。

大家查看代码会发现，代码中实际用的分布是 $Beta(1, 1)$ 。 $Uniform(0, 1)$ 和 $Beta(1, 1)$ 形状相同，而且方便本章后续模拟。

本章代码用到伯努利分布随机数发生器。假设兔子占整体的真实比例为 0.45 (45%)。图 8 (a) 所示为用伯努利随机数发生器产生的随机数，红点 • 代表鸡 (0)，蓝点 ● 代表兔 (1)。

通过图 8 (a) 样本数据做推断便是频率学派的思路。频率学派依靠样本数据，而不引入先验概率(已有知识或主观经验)。当样本数量较大时，频率学派可以做出合理判断；但是，当样本数量很小时，频率学派做出的推断往往不可信。

图 8 (b) 中，从下到上所示为不断抓取动物中鸡、兔各自的比例变化。当动物的数量 n 不断增多时，我们发现比例趋于稳定，并逼近真实值 (0.45)。

图 8 (c) 所示为随着样本数据不断导入，后验概率分布曲线的渐变过程。请大家仔细观察图 8 (c)，看看能不能发现有趣的规律。

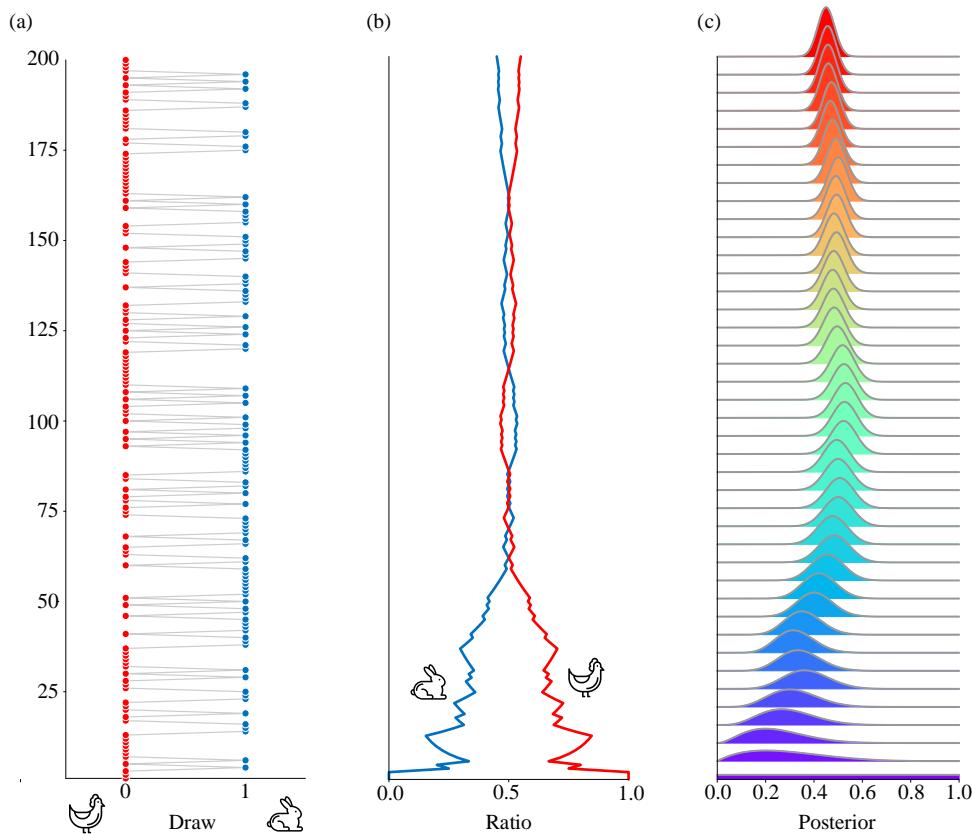
图 8. 某次试验的模拟结果，先验分布为 $\text{Beta}(1, 1)$

图 8(c) 给出的这个过程中，请大家注意两个细节。

第一，后验概率分布 $f_{\theta|x}(\theta|x)$ 曲线不断变的细高，也就是后验标准差不断变小。这是因为样本数据不断增多，大家对鸡兔比例变得越发“确信”。

第二，后验概率分布 $f_{\theta|x}(\theta|x)$ 的最大值，也就是峰值，所在位置逐渐逼近鸡兔的真实比例 0.45。第二点在图 9 中看得更清楚。

图 9(a) 中，先验概率分布为均匀分布，这代表老农对鸡兔比例一无所知。兔子的比例在 0 和 1 之间，任何值皆有可能，而且可能性均等。

图 9(b) 所示为，抓到第一只动物发现是鸡。利用贝叶斯定理，通过图 9(a) 的先验概率（连续均匀分布 $\text{Beta}(1,1)$ ）和样本数据（一只鸡），计算得到图 9(b) 所示的后验概率分布 $\text{Beta}(1,2)$ ，这一过程如图 10 所示。

图 9(b) 这个分布显然认为“农场全是鸡”的可能性更高，但是不排除其他可能。“不排除其他可能”对应图 9(b) 的三角形， θ 在 $[0, 1]$ 区间取值时，后验概率 $f_{\theta|x}(\theta|x)$ 都不为 0。确定的是“农场全是兔”是不可能的，对应概率为 0。

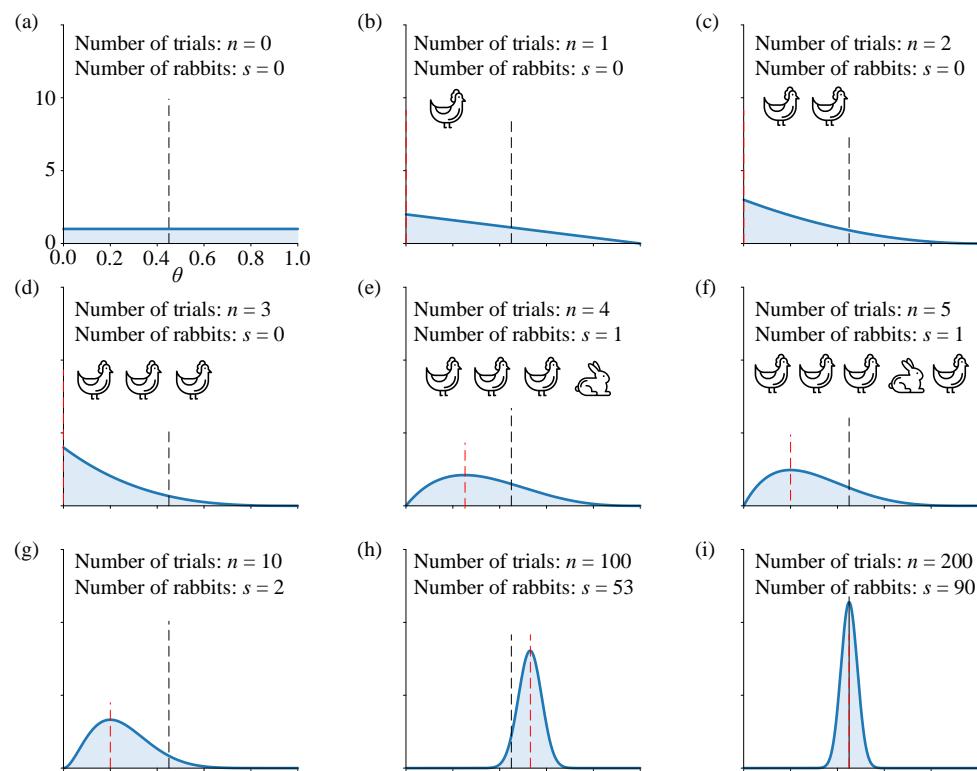


图 9. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(1, 1)

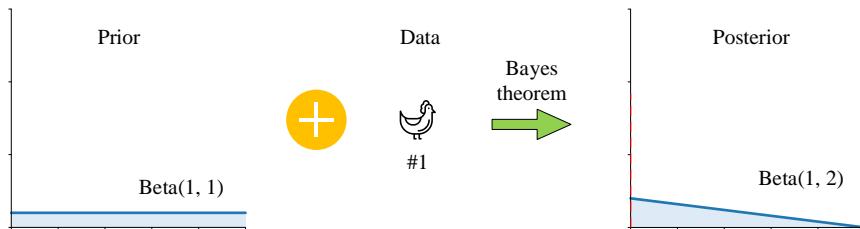


图 10. 不确定鸡兔比例，先验概率 Beta(1, 1) + 一只鸡 (数据) 推导得到后验概率 Beta(1, 2)

抓第二只动物，发现还是鸡。如图 9 (c) 后验概率分布所示，显然农夫心中的天平发生倾斜，认为农场的鸡的比例肯定很高。

获得图 9 (c) 的后验概率分布有两条路径。

第一条如图 11 所示，先验概率 Beta(1, 1) + 两只鸡 (数据) 推导得到后验概率 Beta(1, 3)。

第二条如图 12 所示，更新先验概率 Beta(1, 2) + 第二只鸡 (数据) 推导得到后验概率 Beta(1, 3)。而更新先验概率 Beta(1, 2) 就是图 10 中的后验概率。

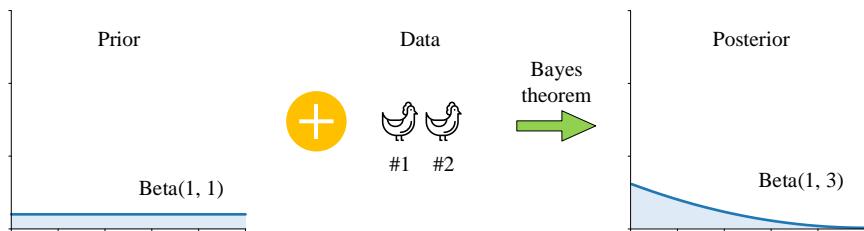


图 11. 第一条路径：先验概率 Beta(1, 1) + 两只鸡 (数据) 推导得到后验概率 Beta(1, 3)

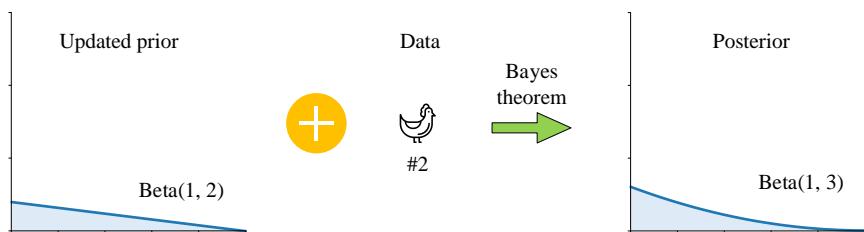


图 12. 第二条路径：更新先验概率 Beta(1, 2) + 第二只鸡 (数据) 推导得到后验概率 Beta(1, 3)

抓第三只动物，竟然还是鸡！如图 9 (d) 所示，农夫心中比例进一步向“鸡”倾斜，但是仍然不能排除其他可能。

理解这步运算则有三条路径！图 13 所示为三条路径中的第一条，请大家自己绘制另外两条。

如果采样此时停止，依照频率派的观点，农场 100%都是鸡。

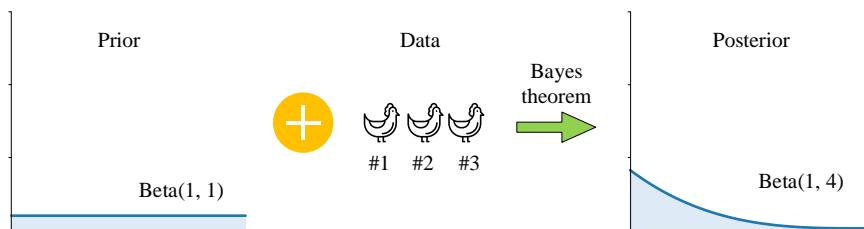


图 13. 先验概率 Beta(1, 1) + 三只鸡 (数据) 推导得到后验概率 Beta(1, 4)

抓第四只动物时，终于抓住一只兔子！农夫才确定农场不都是鸡，确信还是有兔子！观察图 9 (e) 会发现， $\theta = 0$ ，即兔子比例为 0 (或农场全是鸡)，对应的概率密度骤降为 0。

随着抓到的动物不断送来验明正身，农夫的“后验概率”、“先验概率”依次更新。

最终，在抓获的 200 只动物中，有 90 只兔子，也就是说兔子比例 45%。但是观察图 9 (i) 的后验概率曲线，发现 $\theta = 45\%$ 左右的其他 θ 值也不小。

从农夫的视角，农场的鸡兔比例很可能是 45%，但是不排除其他比例的可能性，也就是贝叶斯推断的结论观点。

此外，图9(i)的后验概率的“高矮胖瘦”，也决定了对结论观点的“确信度”。本章后文将展开讲解。

最大化后验概率 MAP

图9中黑色划线为农场兔子的真实比例。

而图9各个子图中红色划线对应的就是后验概率分布的最大值。这便对应贝叶斯推断的优化问题，**最大化后验概率** (Maximum A Posteriori estimation, MAP)：

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} f_{\Theta|x}(\theta|x) \quad (16)$$

将(1)代入上式：

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} \frac{f_{X|\Theta}(x|\theta) f_{\Theta}(\theta)}{\int_{\Theta} f_{X|\Theta}(x|\theta) f_{\Theta}(\theta) d\theta} \quad (17)$$

进一步根据(6)，这个优化问题可以简化为：

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} f_{X|\Theta}(x|\theta) f_{\Theta}(\theta) \quad (18)$$

本书第7章介绍过 Beta(α, β) 分布的众数为：

$$\frac{\alpha-1}{\alpha+\beta-2}, \quad \alpha, \beta > 1 \quad (19)$$

对于本节例子，MAP 的优化解为 Beta($s+1, n-s+1$) 的众数，即概率密度最大值：

$$\hat{\theta}_{\text{MAP}} = \frac{s+1-1}{s+1+n-s+1-2} = \frac{s}{n} \quad (20)$$

兜兜转转，结果这个贝叶斯派 MAP 优化解和频率派 MLE 一致？

MAP 和 MLE 当然不同！

首先，MAP 和 MLE 的优化问题完全不一样，两者分析问题的视角完全不同。回顾 MLE 优化问题：

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} \prod_{i=1}^n f_{X_i}(x_i; \theta) \quad (21)$$

请大家自行对比(16)和(21)。

此外，(20)中这个比例是在先验概率为 Uniform(0, 1) 条件下得到的，下一节大家会看到不同的 MAP 优化结果。

更重要的是，贝叶斯派得到的结论是图9(i)中这个分布。也就是说，最优解虽然在 $\theta = 0.45$ ，但是不排除其他可能。

以图 9 (i) 为例，本例中贝叶斯派得到的参数 θ 为 $\text{Beta}(s + 1, n - s + 1)$ 这个分布。代入具体数据 ($n = 200, s = 90$)，贝叶斯推断的结果为 $\text{Beta}(91, 111)$ ，整个过程如图 14 所示。

图 14 中，先验分布为 $\text{Beta}(1, 1)$ ，括号内的样本数据为 (兔，鸡)，即 (90, 110)，获得的后验概率为 $\text{Beta}(1 + 90, 1 + 110)$ 。 $\text{Beta}(1 + 90, 1 + 110)$ 的标准差可以度量我们对贝叶斯推断结论的确信程度，这是本章最后要讨论的话题之一。

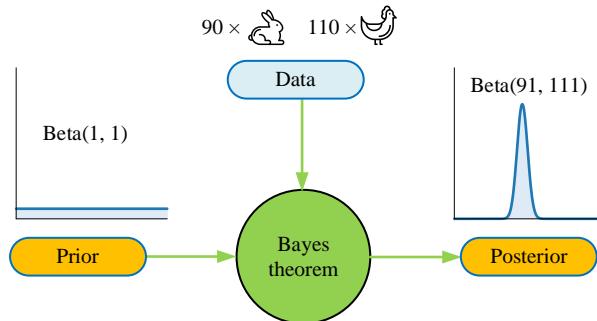


图 14. 先验 $\text{Beta}(1, 1) +$ 样本 $(90, 110) \rightarrow$ 后验 $\text{Beta}(91, 111)$

先验分布的选择和参数的确定代表“经验”，也代表某种“信念”。先验分布的选择和样本数据无关，不需要通过样本数据构造。反过来，观测到的样本数据对先验的选择没有任何影响。

此外，讲解图 12 时，我们看到贝叶斯推断可以采用迭代方式，即后验概率可以成为新样本数据的先验概率。

20.4 走地鸡兔：很可能一半一半

本节我们更换场景，假设农夫认为鸡兔的比例接近 1:1，也就是说，兔子的比例为 50%。但是，农夫对这个比例的确信程度不同。

先验

由于农夫认为鸡兔的比例为 1:1，我们选用 $\text{Beta}(\alpha, \alpha)$ 作为先验分布。 $\text{Beta}(\alpha, \alpha)$ 具体的概率密度函数为：

$$f_{\theta}(\theta) = \frac{1}{B(\alpha, \alpha)} \theta^{\alpha-1} (1-\theta)^{\alpha-1} \quad (22)$$

其中， $\text{Beta}(\alpha, \alpha)$ 为：

$$B(\alpha, \alpha) = \frac{\Gamma(\alpha)\Gamma(\alpha)}{\Gamma(\alpha + \alpha)} \quad (23)$$

再次强调，选取 $\text{Beta}(\alpha, \alpha)$ 和样本无关， $\text{Beta}(\alpha, \alpha)$ 代表事前主观经验。

不同确信程度

图 15 所示为 α 取不同值时 $\text{Beta}(\alpha, \alpha)$ 分布 PDF 图像。

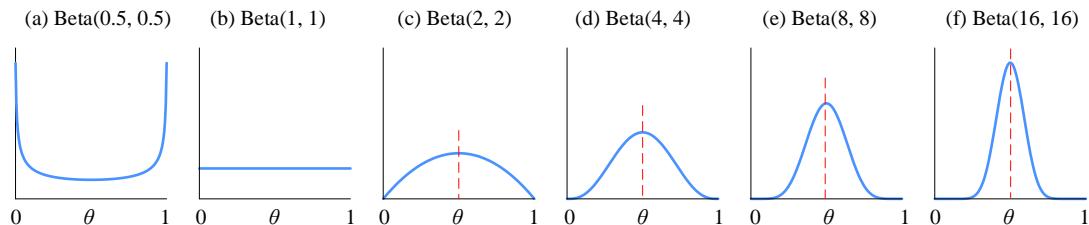


图 15. 五个不同参数 α 取不同值时 $\text{Beta}(\alpha, \alpha)$ 分布 PDF 图像

容易发现 $\text{Beta}(\alpha, \alpha)$ 图像为对称， $\text{Beta}(\alpha, \alpha)$ 的均值和众数为 $1/2$ ，方差为 $1/(8\alpha + 4)$ 。显然，参数 α 小于 1 代表特别“清奇”的观点——农场要么都是鸡、要么都是兔。

α 等于 1 就是本章前文的先验分布为 $\text{Uniform}(0, 1)$ ，即 $\text{Beta}(1, 1)$ ，假设条件。也就是说，当我们事先对比例不持立场，对 $[0, 1]$ 范围内任何一个 θ 值不偏不倚， $\text{Beta}(1, 1)$ 就是最佳的先验分布。

而 α 取不同大于 1 的值时，代表农夫的对鸡兔比例 1:1 的确信程度。

如图 16 所示， α 越大 $\text{Beta}(\alpha, \alpha)$ 的方差越小，这意味着先验分布的图像越窄、越细高，这代表农夫对兔子比例为 50% 这个观点的确信度越高。本章后文会用 Beta 分布的标准差作为“确信程度”的度量，原因是标准差和众数、均值量纲一致。

本节后续的蒙特卡洛模拟中参数 α 的取值分为 2、16 两种情况。 $\alpha = 2$ 代表农夫认为兔子的比例大致 50%，但是确信度不高。 $\alpha = 16$ 则对应农夫认为兔子的比例很可能 50%，但是绝不排除其他比例的可能性，确信度相对高很多。

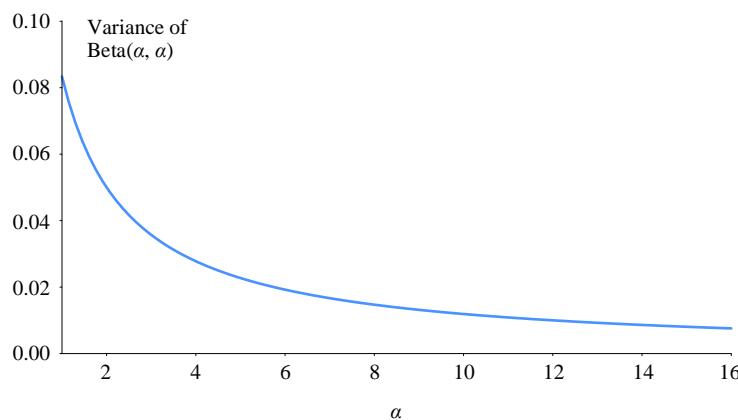


图 16. Beta(α, α) 方差随参数 α 变化

似然

和前文一致，给定 $\Theta = \theta$ 条件下， $X_1, X_2 \dots X_n$ 服从 IID 的伯努利分布 $Bernoulli(\theta)$ ，即：

$$\underbrace{f_{X_i|\Theta}(x_i|\theta)}_{\text{Likelihood}} = \theta^{x_i} (1-\theta)^{1-x_i} \quad (24)$$

似然函数为：

$$f_{X_1, X_2, \dots, X_n|\Theta}(x_1, x_2, \dots, x_n|\theta) = \theta^s (1-\theta)^{n-s} \quad (25)$$

大家可能已经发现，(25) 本质上就是二项分布。二项分布是若干独立的伯努利分布。我们把似然分布记做 $f_{X|\Theta}(x|\theta)$ ：

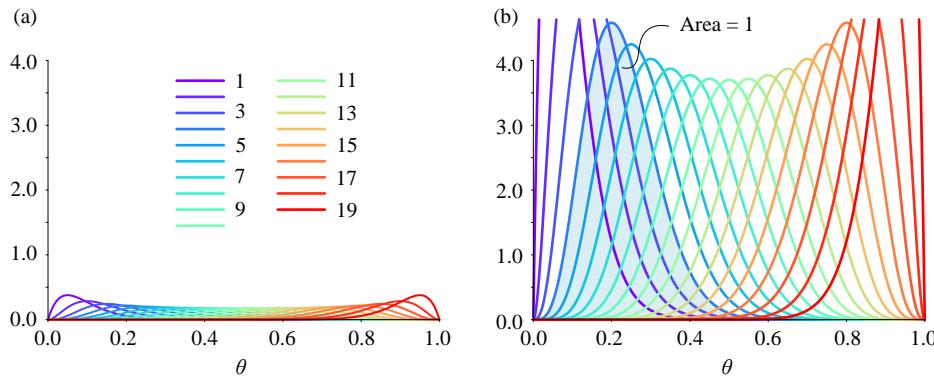
$$f_{X|\Theta}(x|\theta) = C_n^s \cdot \theta^s (1-\theta)^{n-s} \quad (26)$$

C_n^s 和 θ 无关，(46) 和 (26) 成正比关系。也就是说， C_n^s 仅仅提供缩放。

→ 本书第 5 章中，我们这样解读二项分布。给定任意一次试验成功的概率为 θ ，(26) 计算 n 次试验中 s 次成功的概率。对于本例，(26) 的含义是，给定兔子的占比为 θ ， n 只动物中正好有 s 只兔子的概率。

本章中，我们需要换一个视角理解 (26)。它是给定 n 次试验中 s 次成功，而 θ 变化导致概率的变化。而 θ 是在 $(0, 1)$ 区间上连续变化。

图 17 (a) 所示为一组似然分布，其中 $n = 20$ ，这些曲线 s 的取值为 $1 \sim 19$ 整数。 θ 是在 $(0, 1)$ 区间上连续变化。

图 17. 似然分布， $n = 20$

⚠ 注意，似然函数本身是关于 θ 的函数，和先验分布 $\text{Beta}(\alpha, \alpha)$ 中的 α 无关。似然函数值通常是很小的数，所以我们一般会取对数 $\ln()$ 获得对数似然函数。

为了和先验分布、后验分布直接比较，需要归一化(26)：

$$f_{x|\Theta}(x|\theta) = \frac{\overbrace{C_n^s \theta^s (1-\theta)^{n-s}}^{\text{Binomial distribution}}}{\overbrace{C_n^s \int_{\theta} \theta^s (1-\theta)^{n-s} d\theta}^{C_n^s \cdot \text{B}(s+1, n-s+1)}} \quad (27)$$

这样似然函数曲线和横轴围成的面积也是 1。

前文提过，(27) 的分子可以视作二项分布。利用 Beta 函数，(27) 的分母可以进一步化简：

$$C_n^s \int_{\theta} \theta^s (1-\theta)^{n-s} d\theta = C_n^s \cdot \text{B}(s+1, n-s+1) = \frac{n!}{s!(n-s)!} \frac{s!(n-s)!}{(n+1)!} = \frac{1}{n+1} \quad (28)$$

上式就是似然函数的归一化因子。图 17 (b) 所示为归一化后的似然分布。当然我们也可以用数值积分归一化似然函数。

因此，(27) 可以写成：

$$f_{x|\Theta}(x|\theta) = (n+1) \cdot \overbrace{C_n^s \theta^s (1-\theta)^{n-s}}^{\text{Binomial distribution}} \quad (29)$$

在本书第 17 章中，我们知道似然函数的最大值位置为 s/n ，也就是最大似然估计 MLE 的解，具体位置如图 18 所示。注意图 18 中， s 为 $0 \sim 20$ 的整数。

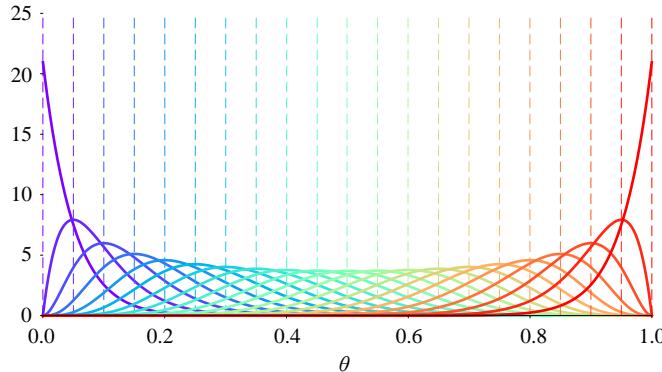


图 18. 似然分布和 MLE 优化解的位置， $n = 20$

再换个视角，看到(25)这种形式，大家是否立刻想到，这不正是一个 Beta 分布！缺的就是归一化系数！补齐这个归一化系数，我们便得到 $\text{Beta}(s+1, n-s+1)$ 分布：

$$\frac{\Gamma(s+1+n-s+1)}{\Gamma(s+1)\Gamma(n-s+1)} \theta^{s+1-1} (1-\theta)^{n-s+1-1} = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} \theta^{s+1-1} (1-\theta)^{n-s+1-1} \quad (30)$$

而 $\text{Beta}(s+1, n-s+1)$ 分布的众数位置为：

$$\frac{s+1-1}{s+1+n-s+1-2} = \frac{s}{n} \quad (31)$$

这和之前的结论一致。请大家自己绘制 $n=20$ 、 s 为 $0 \sim 20$ 整数时， $\text{Beta}(s+1, n-s+1)$ 的 PDF 曲线，并和图 18 比较。

回看 (14)，本节的似然分布 $\text{Beta}(s+1, n-s+1)$ 相当于对鸡兔比例“不持立场”，一切均以客观样本数据为准。

再换个角度来看，上述讨论似乎说明，贝叶斯推断“包含了”频率推断。MLE 是 MAP 的特例（无信息先验）。

先验 vs 似然

图 19 中灰色曲线对应“归一化”的似然分布 $f_{X|\Theta}(x|\theta)$ ，它相当于 $\text{Beta}(s+1, n-s+1)$ 。灰色划线对应 MLE 的解， $f_{X|\Theta}(x|\theta)$ 的最大值。

图 19 中粉色曲线对应 $f_\Theta(\theta)$ ，即 $\text{Beta}(\alpha, \alpha)$ 。如 (22) 所示， $f_\Theta(\theta)$ 和 α 有关； α 越大， $f_\Theta(\theta)$ 曲线越细高。 $f_\Theta(\theta)$ 曲线的最大值是 $\text{Beta}(\alpha, \alpha)$ 的众数， $\theta = 1/2$ 。

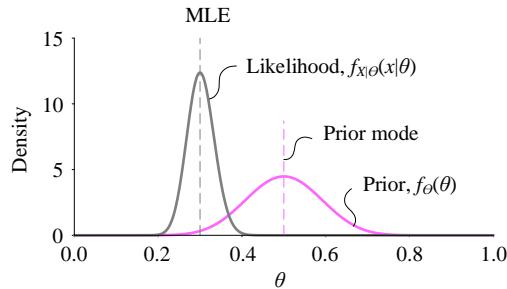


图 19. 对比先验分布、似然分布， $\alpha = 16$

联合

联合分布为：

$$\begin{aligned} f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta) &= \underbrace{f_{X_1, X_2, \dots, X_n | \Theta}(x_1, x_2, \dots, x_n | \theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} \\ &= \theta^s (1-\theta)^{n-s} \frac{1}{B(\alpha, \alpha)} \theta^{\alpha-1} (1-\theta)^{\alpha-1} \\ &= \frac{1}{B(\alpha, \alpha)} \theta^{s+\alpha-1} (1-\theta)^{n-s+\alpha-1} \end{aligned} \quad (32)$$

证据

证据因子 $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ 可以通过 $f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta)$ 对 θ “偏积分”得到：

$$\begin{aligned} f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= \int_{\theta} f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta) d\theta \\ &= \frac{1}{B(\alpha, \alpha)} \int_{\theta} \theta^{s+\alpha-1} (1-\theta)^{n-s+\alpha-1} d\theta \\ &= \frac{B(s+\alpha, n-s+\alpha)}{B(\alpha, \alpha)} \end{aligned} \quad (33)$$

后验

在 $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ 条件下， Θ 的后验分布为：

$$\begin{aligned} f_{\Theta|X_1, X_2, \dots, X_n}(\theta|x_1, x_2, \dots, x_n) &= \frac{f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta)}{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)} \\ &= \frac{\frac{1}{B(\alpha, \alpha)} \theta^{s+\alpha-1} (1-\theta)^{n-s+\alpha-1}}{\frac{B(s+\alpha, n-s+\alpha)}{B(\alpha, \alpha)}} = \frac{\theta^{s+\alpha-1} (1-\theta)^{n-s+\alpha-1}}{B(s+\alpha, n-s+\alpha)} \end{aligned} \quad (34)$$

上式对应 $Beta(s + \alpha, n - s + \alpha)$ 分布。

幸运的是，我们实际上“避开”(33)这个复杂积分。但是，并不是所有情况都存在积分的**闭式解**(closed form solution)，也叫**解析解**(analytical solution)。

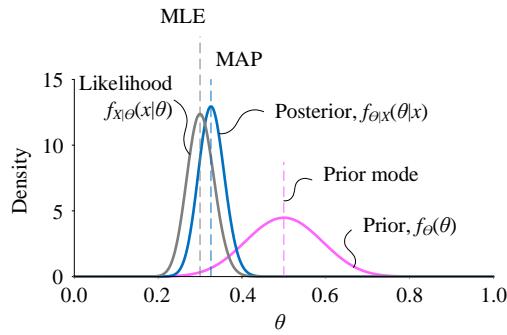


本书第 22 章将介绍蒙特卡洛模拟方式近似获得后验分布。

先验 vs 似然 vs 后验

图 20 对比对比先验分布 $Beta(\alpha, \alpha)$ 、似然分布 $Beta(s+1, n-s+1)$ 、后验分布 $Beta(s+\alpha, n-s+\alpha)$ 。

比较这三个分布，直觉告诉我们后验分布 $Beta(s+\alpha, n-s+\alpha)$ 好像是先验分布 $Beta(\alpha, \alpha)$ 、似然分布 $Beta(s+1, n-s+1)$ 的某种“糅合”！本章最后会继续这个思路探讨贝叶斯推断。

图 20. 对比先验分布、似然分布、后验分布， $\alpha = 16$

正比关系

类似(15)，后验概率存在如下正比关系：

$$f_{\Theta|X_1, X_2, \dots, X_n}(\theta | x_1, x_2, \dots, x_n) \propto f_{X_1, X_2, \dots, X_n | \Theta}(x_1, x_2, \dots, x_n | \theta) f_\Theta(\theta) = \theta^{s+\alpha-1} (1-\theta)^{n-s+\alpha-1} \quad (35)$$

蒙特卡罗模拟：确信度不高

前文提到，农夫认为农场兔子的比例大致为 50%，因此我们选择 Beta(α, α) 作为先验概率分布。下面的蒙特卡罗模拟中，我们设定 $\alpha = 2$ 。

图 21 (a) 所示为伯努利随机数发生器产生的随机数。和前文一样，0 代表鸡，1 代表兔。不同的是，我们设定兔子的真实比例为 0.3。

如图 21 (b) 所示，随着样本数 n 增大，鸡兔的比例趋于稳定。

图 21 (c) 所示为后验概率分布随着 n 的变化。自下而上，后验概率曲线从平缓逐渐过渡到细高，这代表确信度不断升高。

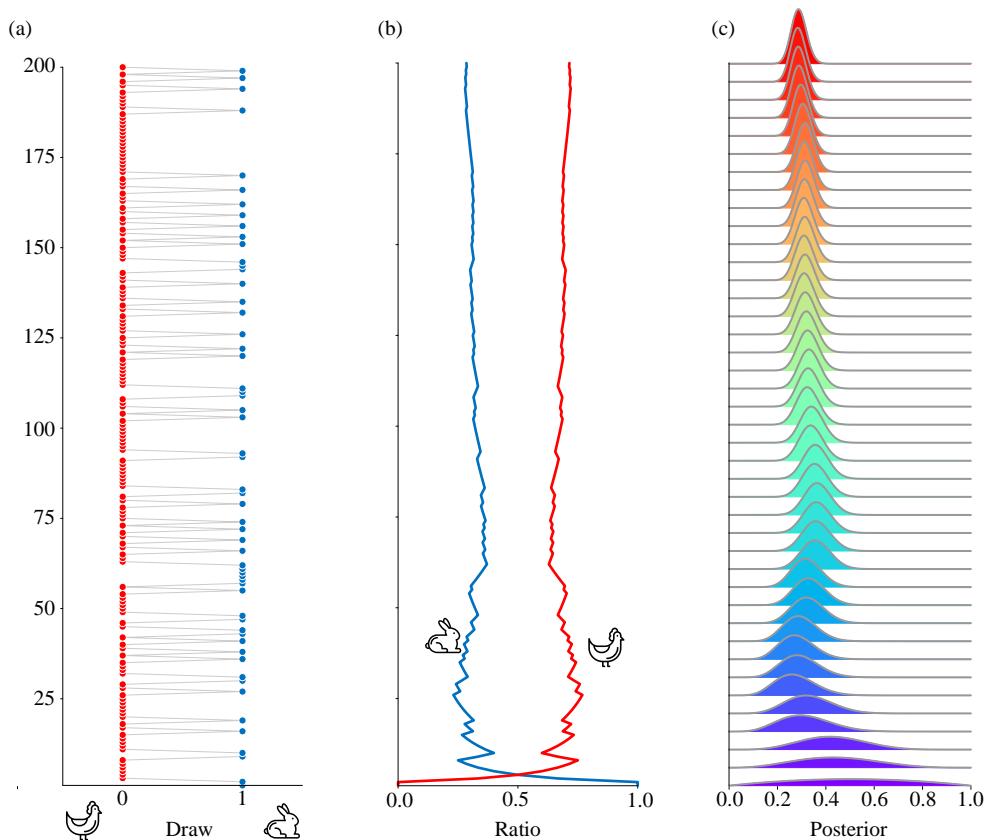
图 21. 某次试验的模拟结果，先验分布为 $\text{Beta}(2, 2)$

图 22 所示为九张不同节点的后验概率分布曲线快照。

图 22 (a) 代表农夫最初的先验概率 $\text{Beta}(2, 2)$ 。 $\text{Beta}(2, 2)$ 曲线关于 $\theta = 0.5$ 对称，并在 $\theta = 0.5$ 取得最大值。 $\text{Beta}(2, 2)$ 很平缓，这代表农夫对 50% 的比例不够确信。

抓到第一只动物是兔子，这个样本导致图 22 (b) 中后验概率最大值向右移动。请大家自己写出后验 Beta 分布的参数。

抓到的第二只动物还是兔子，后验概率最大值进一步向右移动，具体如图 22 (c) 所示。

第三只动物是鸡，后验概率最大值所在位置向左移动了一点。

请大家自行分析图 22 剩下几幅子图，注意后验概率形状、最大值位置变化。

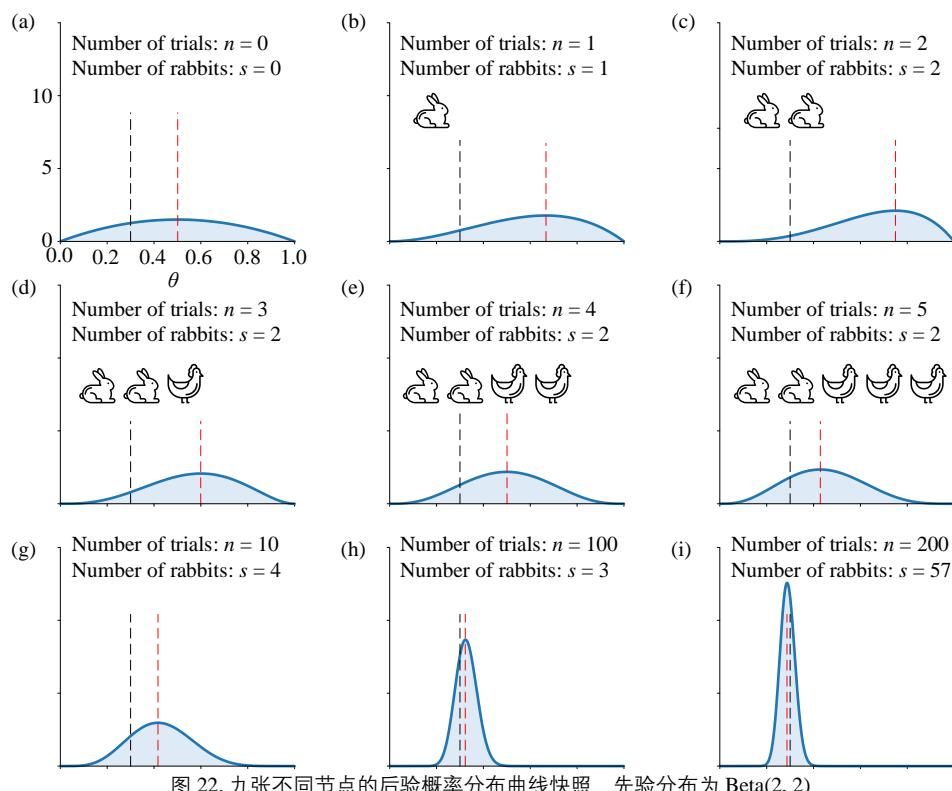


图 22. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(2, 2)

蒙特卡罗模拟：确信度很高

$\alpha = 16$ 则对应农夫认为兔子的比例很可能 50%，但是绝不排除其他比例的可能性，确信度相对高很多。请大家对比前文蒙特卡洛模拟结果，自行分析图 23 和图 24。

强烈建议大家把图 24 每幅子图的 Beta 分布的参数写出来。

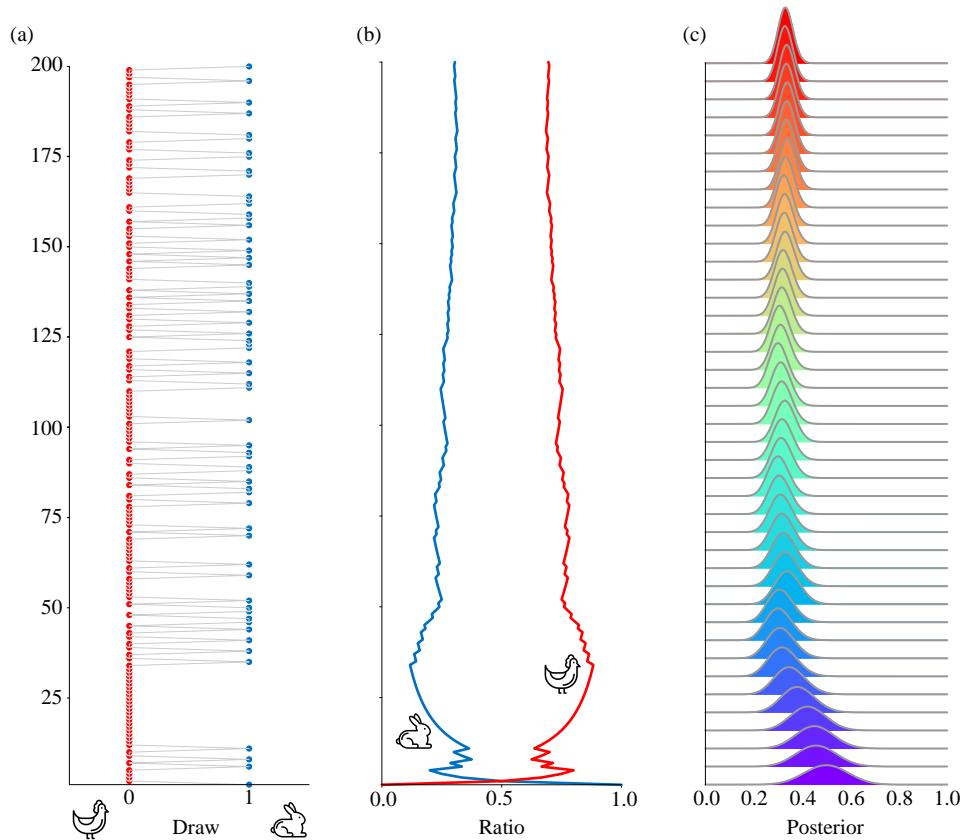


图 23. 某次试验的模拟结果，先验分布为 Beta(16, 16)

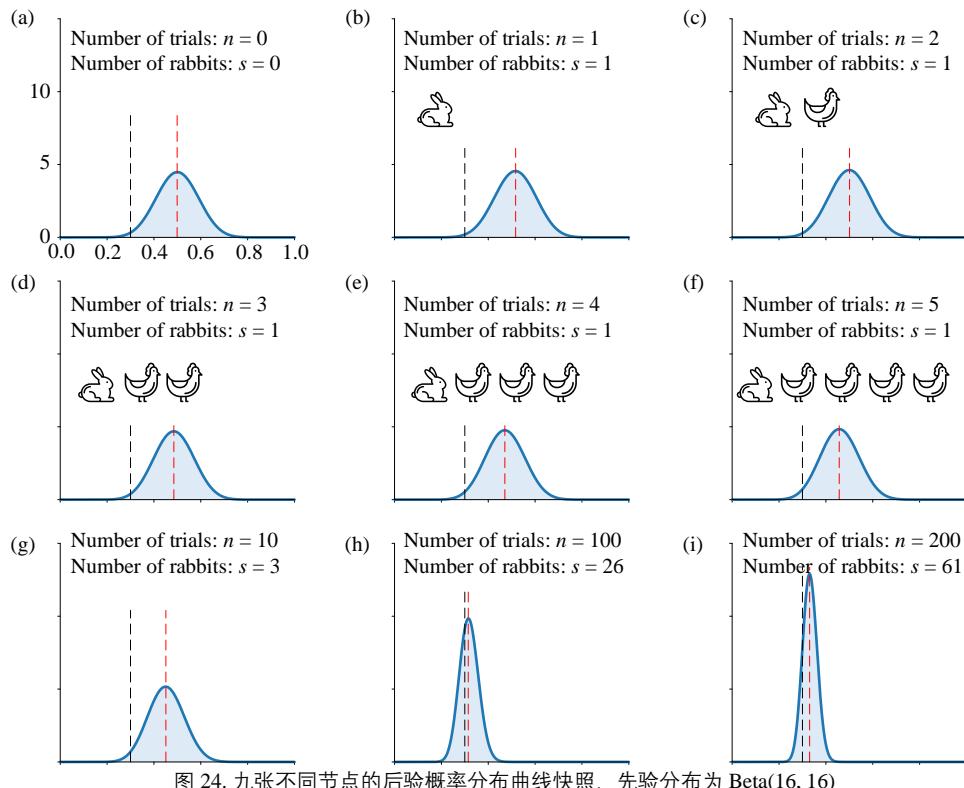


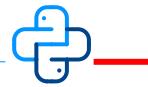
图 24. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(16, 16)

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com



代码 Bk5_Ch20_01.py 完成本章前文蒙特卡洛模拟和可视化。

最大后验 MAP

Beta($s + \alpha, n - s + \alpha$) 的众数，即 MAP 的优化解，为：

$$\hat{\theta}_{\text{MAP}} = \frac{s + \alpha - 1}{n + 2\alpha - 2} \quad (36)$$

特别地，当 $\alpha = 1$ 时，MAP 和 MLE 的解相同，即：

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}} = \frac{s}{n} \quad (37)$$

图 25 对比 α 取不同值时先验分布、似然分布、后验分布。先验分布 Beta(α, α) 中 α 越大，代表主观经验越发“先入为主”，对贝叶斯推断最终结果越强。表现在图 25 中就是，随着 α 增大，似然分布和后验分布差异越大，MAP 优化解越发偏离 MLE 优化解。

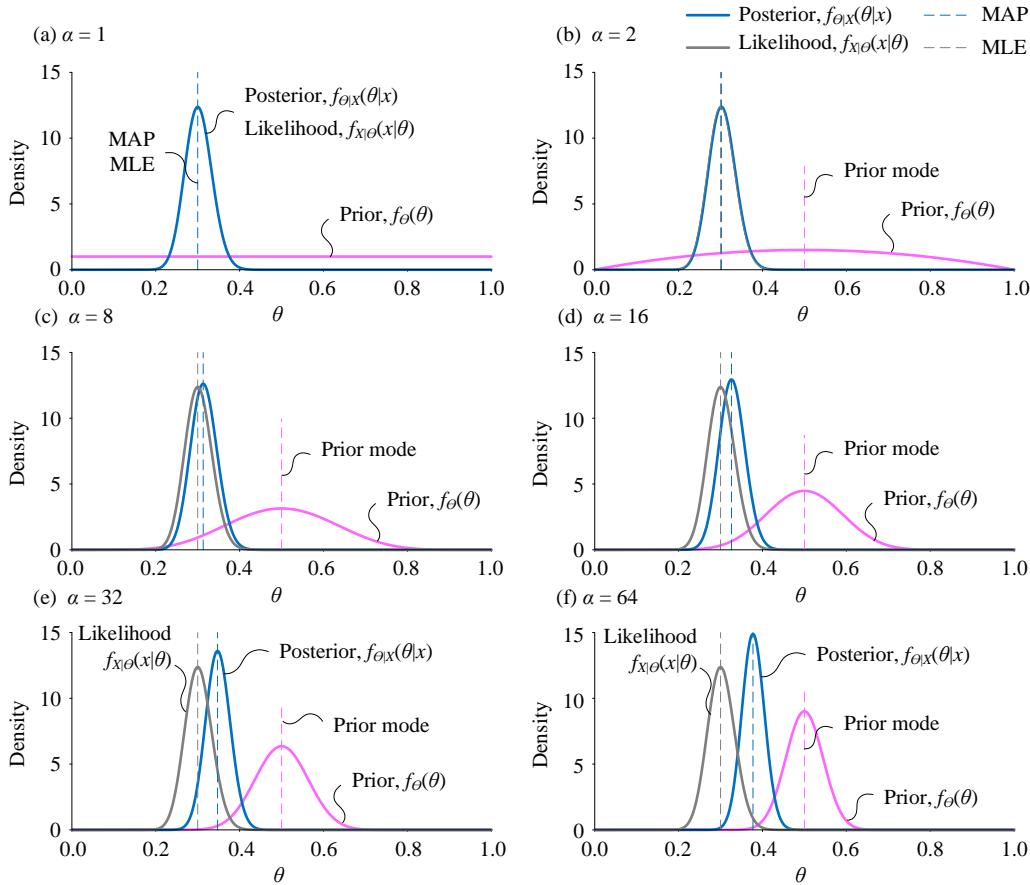
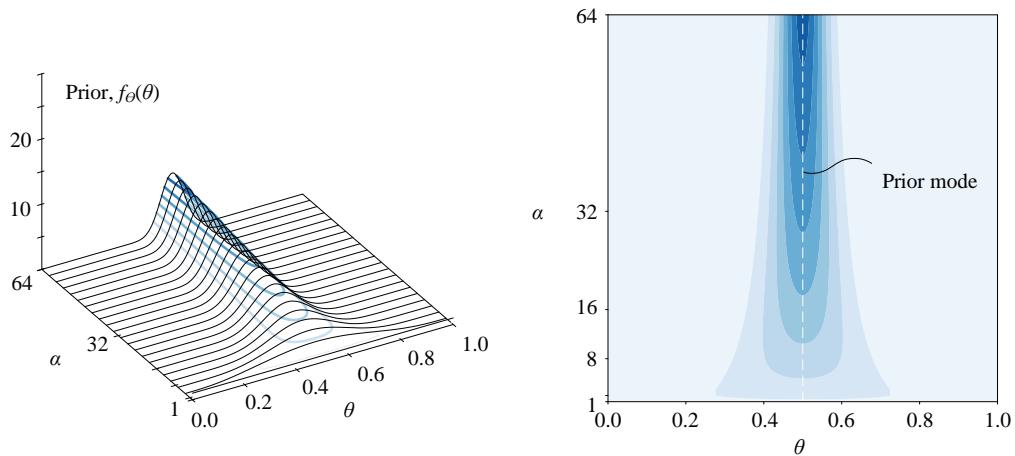
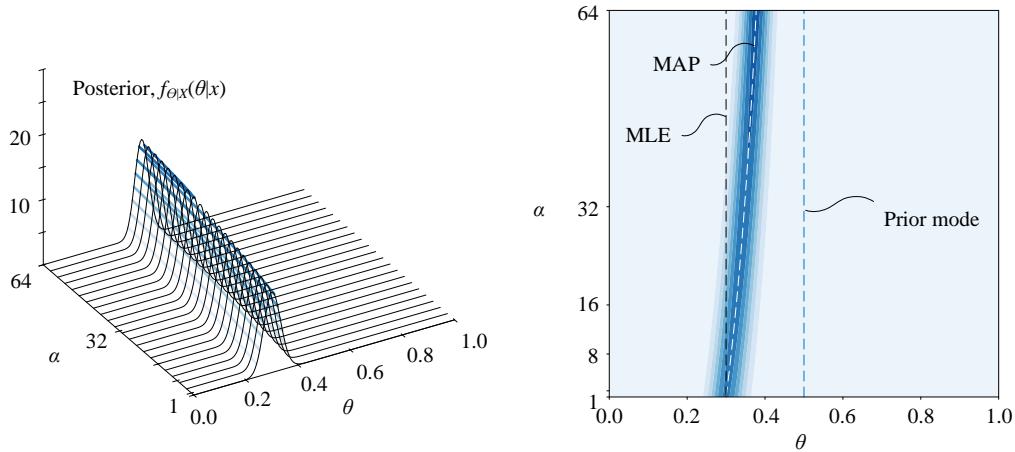
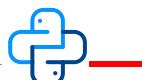


图 25. 对比先验分布、似然分布、后验分布， α 取不同值时

图 26 和图 27 以另外一种可视化方案对比 α 取不同值时先验分布对后验分布的影响。图 26. 先验分布, α 取不同值时图 27. 后验分布, α 取不同值时

代码 Bk5_Ch021_02.py 绘制图 25、图 26、图 27。

20.5 走地鸡兔：更一般的情况

有了前文的两个例子，下面我们看一下更为一般的情况。

先验

选用 $\text{Beta}(\alpha, \beta)$ 作为先验分布。 $\text{Beta}(\alpha, \beta)$ 具体的概率密度函数为：

$$f_{\Theta}(\theta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \quad (38)$$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

先验分布 $\text{Beta}(\alpha, \beta)$ 的众数为：

$$\frac{\alpha-1}{\alpha+\beta-2}, \quad \alpha, \beta > 1 \quad (39)$$

其他比例

举个例子，假设农夫认为兔子比例为 $1/3$ ，则：

$$\frac{\alpha-1}{\alpha+\beta-2} = \frac{1}{3} \quad (40)$$

即 α 和 β 关系为：

$$\beta = 2\alpha - 1 \quad (41)$$

图 28 所示为 α 和 β 取不同值时 $\text{Beta}(\alpha, \beta)$ 分布 PDF 图像。这些图像有一个共同特点，众数都是 $1/3$ 。

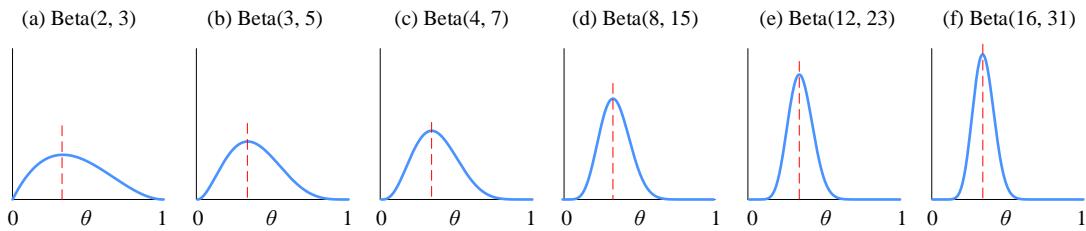


图 28. 五个不同 $\text{Beta}(\alpha, \beta)$ 分布 PDF 图像，众数都是 $1/3$

如果农夫认为兔子比例为 $1/4$ ，则：

$$\frac{\alpha-1}{\alpha+\beta-2} = \frac{1}{4} \quad (42)$$

即 α 和 β 关系为：

$$\beta = 3\alpha - 2 \quad (43)$$

满足上式条件下，当 α 不断增大，兔子的比例虽然还是 $1/4$ ，但是如图 29 所示，先验分布变得越发细高，这代表着确信程度提高，“信念”增强。

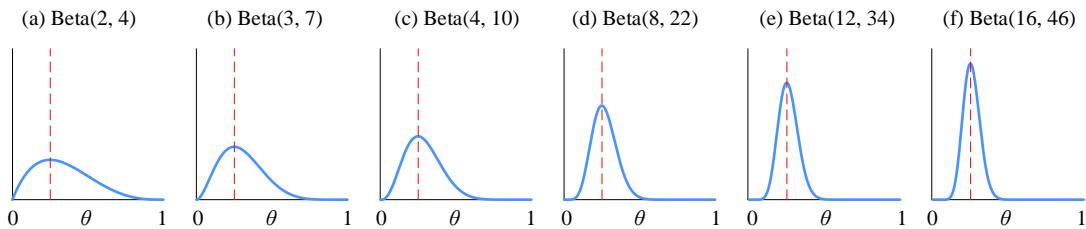


图 29. 五个不同 $\text{Beta}(\alpha, \beta)$ 先验分布 PDF 图像，众数都是 $1/4$

确信程度

我们可以用 $\text{Beta}(\alpha, \beta)$ 分布的标准差量化所谓“确信程度”。

$\text{Beta}(\alpha, \beta)$ 的标准差为：

$$\text{std}(X) = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \quad (44)$$

如果 α, β 满足 (43) 等式， $\text{Beta}(\alpha, \beta)$ 的标准差随 α 变化如图 30 所示。更准确地说，随着标准差减小，对比例的“怀疑程度”不断减小。

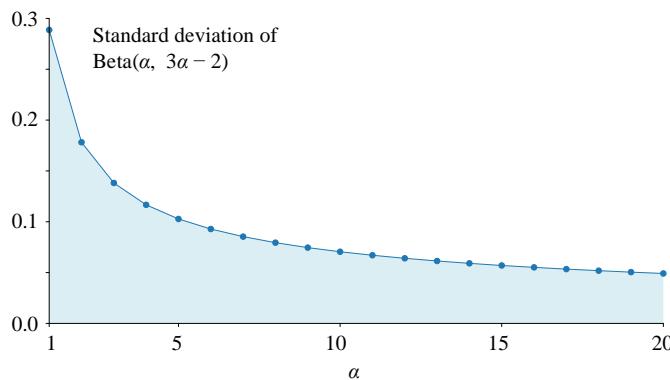


图 30. 随着 α 增大，“怀疑程度”不断减小

换一个方式，为了方便和下一章的 Dirichlet 分布对照，令 $\alpha_0 = \alpha + \beta$ ， $\text{Beta}(\alpha, \beta)$ 的均方差可以进一步写成：

$$\text{std}(X) = \sqrt{\frac{\alpha/\alpha_0(1-\alpha/\alpha_0)}{\alpha_0+1}} \quad (45)$$

α/α_0 也可以看做兔子的比例。不同的是， α/α_0 代表 $\text{Beta}(\alpha, \beta)$ 的期望（均值），不是众数。下一章会比较 Beta 分布的期望和均值。

图 31 所示一组图像代表比例和确信度同时变化。

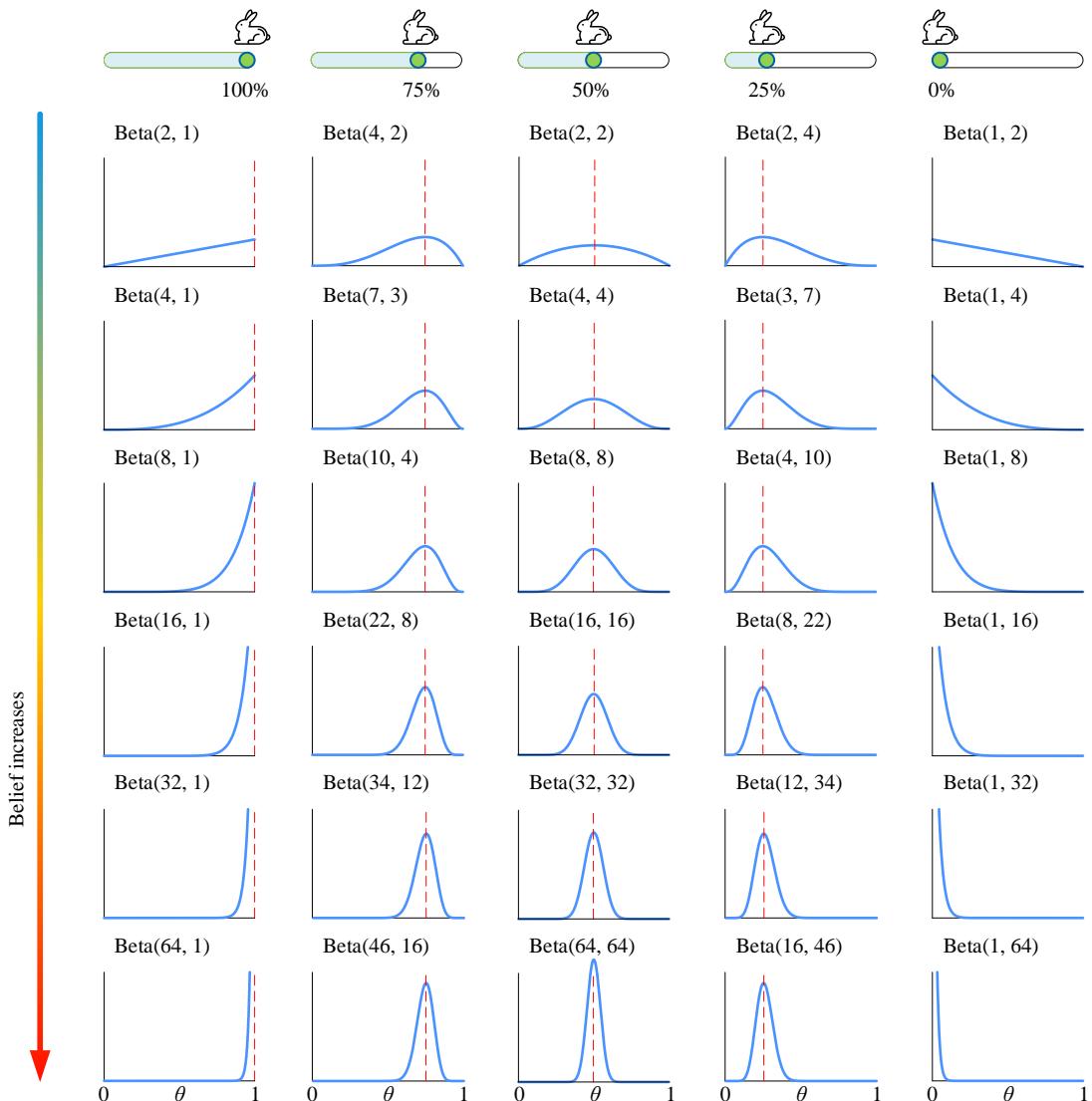


图 31. 比例和确信程度同时变化

似然

和前文一致，似然函数为：

$$f_{X_1, X_2, \dots, X_n | \Theta}(x_1, x_2, \dots, x_n | \theta) = \theta^s (1-\theta)^{n-s} \quad (46)$$

本章前文介绍过，似然函数可以看成 IID 伯努利分布、二项分布，甚至用 Beta 分布代替。

联合

因此，联合分布为：

$$\begin{aligned}
 f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta) &= \underbrace{f_{X_1, X_2, \dots, X_n | \Theta}(x_1, x_2, \dots, x_n | \theta)}_{\text{Likelihood}} \underbrace{f_{\Theta}(\theta)}_{\text{Prior}} \\
 &= \theta^s (1-\theta)^{n-s} \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\
 &= \frac{1}{B(\alpha, \beta)} \theta^{s+\alpha-1} (1-\theta)^{n-s+\beta-1}
 \end{aligned} \tag{47}$$

证据

证据因子 $f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)$ 可以通过 $f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta)$ 对 θ “偏积分”得到：

$$\begin{aligned}
 f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) &= \int_{\theta} f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta) d\theta \\
 &= \frac{1}{B(\alpha, \beta)} \int_{\theta} \theta^{s+\alpha-1} (1-\theta)^{n-s+\beta-1} d\theta \\
 &= \frac{B(s+\alpha, n-s+\beta)}{B(\alpha, \beta)}
 \end{aligned} \tag{48}$$

后验

在 $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ 条件下， Θ 的后验分布为：

$$\begin{aligned}
 f_{\Theta | X_1, X_2, \dots, X_n}(\theta | x_1, x_2, \dots, x_n) &= \frac{f_{X_1, X_2, \dots, X_n, \Theta}(x_1, x_2, \dots, x_n, \theta)}{f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n)} \\
 &= \frac{\frac{1}{B(\alpha, \beta)} \theta^{s+\alpha-1} (1-\theta)^{n-s+\beta-1}}{\frac{B(s+\alpha, n-s+\beta)}{B(\alpha, \beta)}} = \frac{\theta^{s+\alpha-1} (1-\theta)^{n-s+\beta-1}}{B(s+\alpha, n-s+\beta)}
 \end{aligned} \tag{49}$$

上式对应 $Beta(s + \alpha, n - s + \beta)$ 分布。

看到这里，大家肯定会想我们是幸运的，因为我们再次成功地避开了 (48) 这个复杂的积分。而这绝不是巧合！在贝叶斯统计中，如果后验分布 $Beta(s + \alpha, n - s + \beta)$ 与先验分布 $Beta(\alpha, \beta)$ 属于同类，则先验分布与后验分布被称为**共轭分布** (conjugate distribution 或 conjugate pair)，而先验分布被称为似然函数的**共轭先验** (conjugate prior)。



下一章还会探讨共轭分布这一话题。

贝叶斯收缩

$Beta(s + \alpha, n - s + \beta)$ 的众数为：

$$\frac{s + \alpha - 1}{n + \alpha + \beta - 2} \quad (50)$$

我们可以把上式写成两个部分：

$$\begin{aligned} \frac{s + \alpha - 1}{n + \alpha + \beta - 2} &= \frac{\alpha - 1}{n + \alpha + \beta - 2} + \frac{s}{n + \alpha + \beta - 2} \\ &= \frac{\alpha + \beta - 2}{n + \alpha + \beta - 2} \times \underbrace{\frac{\alpha - 1}{\alpha + \beta - 2}}_{\text{Prior mode}} + \frac{n}{n + \alpha + \beta - 2} \times \underbrace{\frac{s}{n}}_{\text{Sample mean}} \end{aligned} \quad (51)$$

定义权重：

$$\begin{aligned} w &= \frac{\alpha + \beta - 2}{n + \alpha + \beta - 2} \\ 1 - w &= \frac{n}{n + \alpha + \beta - 2} \end{aligned} \quad (52)$$

(51) 可以写成：

$$\frac{s + \alpha - 1}{n + \alpha + \beta - 2} = w \times \underbrace{\frac{\alpha - 1}{\alpha + \beta - 2}}_{\text{Prior mode}} + (1 - w) \times \underbrace{\frac{s}{n}}_{\text{Sample mean}} \quad (53)$$

随着 n 不断增大， w 趋向于 0，而 $1 - w$ 趋向于 1。也就是说，随着样本数据量不断增多，先验的影响力不断减小。 $n \rightarrow \infty$ 时，MAP 和 MLE 的结果趋同。

相反，当 n 较小的时候，特别是当 α 和 β 比较大，则先验的影响力很大，MAP 的结果向先验均值“收缩”。这种效果常被称作**贝叶斯收缩** (Bayes shrinkage)。

贝叶斯收缩也可以从期望角度理解。Beta($s + \alpha, n - s + \beta$) 的期望也可以写成两部分：

$$\begin{aligned} \frac{s + \alpha}{n + \alpha + \beta} &= \frac{\alpha}{n + \alpha + \beta} + \frac{s}{n + \alpha + \beta} \\ &= \frac{\alpha + \beta}{n + \alpha + \beta} \times \frac{\alpha}{\alpha + \beta} + \frac{n}{n + \alpha + \beta} \times \underbrace{\frac{s}{n}}_{\text{Sample mean}} \end{aligned} \quad (54)$$

从贝叶斯收缩角度，让我们再回过头来看本节上述结果。

首先，换个视角理解先验分布 Beta(α, β) 中的 α 和 β 。

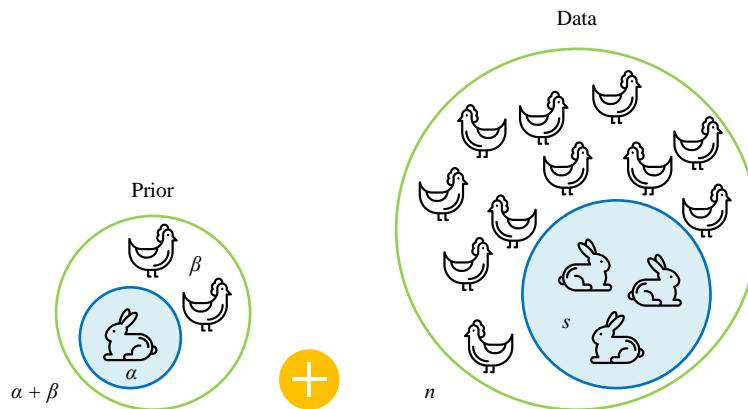


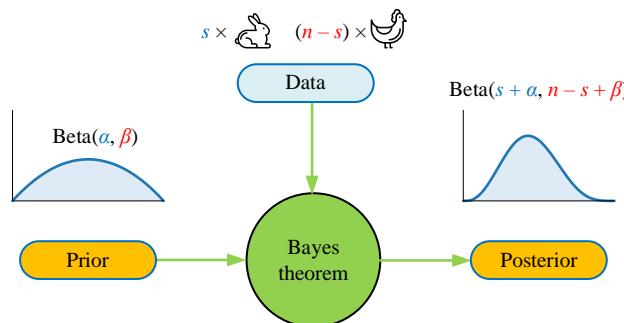
图 32. “混合”先验、样本数据

先验分布中的 α 和 β 之和可以看做“先验”动物总数。即没有数据时，根据先验经验，农夫认为农场动物总数为 $\alpha + \beta$ ，其中兔子的比例为 $\alpha/(\alpha + \beta)$ 。

样本数据中， s 代表 n 只动物中兔子的数量， $n - s$ 代表鸡的数量，兔子比例为 s/n 。

而 (54) 就可以简单理解成“先验 + 数据”融合得到“后验”。

后验分布 $\text{Beta}(s + \alpha, n - s + \beta)$ 则代表“先验 $\text{Beta}(\alpha, \beta)$ + 数据 $(s, n - s)$ ”。兔子 α 从增加到 $s + \alpha$ ，鸡从 β 增加到 $n - s + \beta$ 。

图 33. 先验 $\text{Beta}(\alpha, \beta)$ + 样本 $(s, n - s)$ → 后验 $\text{Beta}(s + \alpha, n - s + \beta)$

当然， α 和 β 越大，先验的“主观”影响力越大。但是随着样本数量不断增大，先验的影响力逐步下降。当样本数量趋近无穷时，先验不再有任何影响力，MAP 优化解趋向于 MLE 优化解。

换个角度，当我们对参数先验知识模糊不清时， $\text{Beta}(1, 1)$ 并非唯一选择。任何 α 和 β 较小的 Beta 分布都可以。因为随着样本数量不断增大，先验分布的较小参数对后验影响微乎其微。



有趣的是，贝叶斯推断所体现出来的“学习过程”和人类认知过程极为相似。贝叶斯推断的优点在于其能够利用先验信息和后验概率，通过不断更新来获得更准确的估计结果。

总结来说，贝叶斯推断的过程包括以下几个步骤：1) 确定模型和参数空间，建立参数的先验分布；2) 收集数据；3) 根据样本数据，计算似然函数；4) 利用贝叶斯定理，将似然函数与先验概率相结合，计算后验概率；5) 根据后验概率，更新先验概率，得到更准确的参数估计。

本章透过二项比例的贝叶斯推断，以 Beta 分布为先验，以伯努利分布或二项分布作为似然分布，讨论不同参数对贝叶斯推断结果的影响。

请大家格外注意，这仅仅是众多贝叶斯推断中较为简单的一种。虽然以管窥豹，希望大家能通过本章例子理解贝叶斯推断背后的思想，以及整条技术路线。此外，本章和下两章共用一幅思维导图。

本章农场仅有鸡、兔，即二元。下一章中，农场又来了猪，贝叶斯推断变成了三元，进一步“升维”。先验分布则变成了 Dirichlet 分布，似然分布为多项分布。

21

Dive into Bayesian Inference

贝叶斯推断进阶

属于同类的后验分布与先验分布叫共轭分布



生活中没有什么是可怕的，它们只是需要被理解。现在是了解更多的时候了，这样我们就可以减少恐惧。

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

—— 玛丽·居里 (Marie Curie) | 波兰裔法国籍物理学家、化学家 | 1867 ~ 1934



- ◀ matplotlib.pyplot.axvline() 绘制竖直线
- ◀ matplotlib.pyplot.fill_between() 区域填充颜色
- ◀ numpy.cumsum() 累加
- ◀ scipy.stats.bernoulli.rvs() 满足伯努利分布的随机数
- ◀ scipy.stats.beta() Beta 分布 scipy.stats.beta() Beta 分布
- ◀ scipy.stats.beta.pdf() Beta 分布概率密度函数
- ◀ scipy.stats.dirichlet() Dirichlet 分布
- ◀ scipy.stats.dirichlet.pdf() Dirichlet 分布概率密度函数

21.1 除了鸡兔，农场发现了猪

鸡、兔、猪同笼

在确定农场走地鸡兔比例时，农夫发现农场还有大量的“走地”猪！

为了搞清楚农场鸡、兔、猪比例，农夫决定随机抓 n 只动物。 $X_1, X_2 \dots X_n$ 为每次抓取动物的结果。 X_i 的样本空间为 $\{0, 1, 2\}$ ，其中 0 代表鸡，1 代表兔，2 代表猪。和上一章一样，忽略抓取动物对农场整体动物总体比例的影响。

下面我们采用和上一章完全一样，以“先验 \rightarrow 似然 \rightarrow 后验”的思路来进行贝叶斯推断。

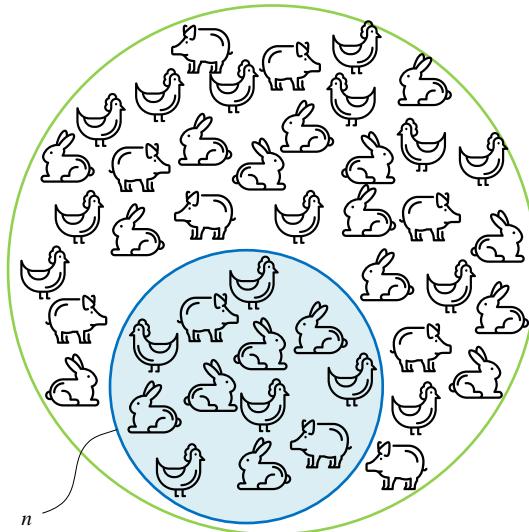


图 1. 农场有数不清的散养鸡兔猪

先验分布

在出现样本数据之前，先验分布代表我们对模型参数的既有“知识”，主观“经验”。

$\theta_1, \theta_2, \theta_3$ 分别为农场中鸡、兔、猪的比例， $\theta_1, \theta_2, \theta_3$ 的取值范围都是 $[0, 1]$ 。鸡兔猪比例之和为 1，即 $\theta_1, \theta_2, \theta_3$ 满足如下等式：

$$\theta_1 + \theta_2 + \theta_3 = 1 \quad (1)$$

我们把 $\theta_1, \theta_2, \theta_3$ 写成一个向量 $\boldsymbol{\theta}$ 。

上一章中，我们采用 Beta 分布作为先验分布。这一章，鸡兔猪问题中 $\boldsymbol{\theta} \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ ：

$$f_{\Theta}(\boldsymbol{\theta}) = \frac{1}{B(\alpha_1, \alpha_2, \alpha_3)} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \theta_3^{\alpha_3-1} \quad (2)$$

$B(\boldsymbol{\alpha})$ 起到“归一化”作用，具体定义为：

$$B(\alpha_1, \alpha_2, \alpha_3) = \frac{\prod_{i=1}^3 \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^3 \alpha_i\right)} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)} \quad (3)$$



本书第 7 章提过，Dirichlet 分布也叫狄利克雷分布，它本质上是多元 Beta 分布。或者说，Beta 分布是特殊的 Dirichlet 分布。

我们也可以把 $\text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ 写成 $\text{Dir}(\boldsymbol{\alpha})$ 。

先验分布位置

通过上一章学习我们知道，对于一个先验分布，常用众数、期望（均值）描述它的位置。

对于 $\text{Dir}(\boldsymbol{\alpha})$ ， X_i 的众数为：

$$\frac{\alpha_i - 1}{\sum_{k=1}^K \alpha_k - K} = \frac{\alpha_i - 1}{\alpha_0 - K}, \quad \alpha_i > 1 \quad (4)$$

这是先验初始比例所在位置，也是 MAP 的位置。其中， $\alpha_0 = \sum_{k=1}^K \alpha_k$ 。

特别地如果 $\alpha_1 = \alpha_2 = \dots = \alpha_K$ ， X_i 的众数为：

$$\frac{\alpha_i - 1}{\alpha_0 - K} = \frac{1}{K}, \quad \alpha_i > 1 \quad (5)$$

对于 $\text{Dir}(\boldsymbol{\alpha})$ ， X_i 的期望为：

$$\frac{\alpha_i}{\sum_{k=1}^K \alpha_k} = \frac{\alpha_i}{\alpha_0} \quad (6)$$

此外，大家可能会想到**中位数** (median)，也就是百分位 50-50 的位置。本章马上比较众数、期望、中位数。

似然分布

在贝叶斯推断中，我们用似然分布整合样本数据，并描述样本分布。

⚠ 注意，似然函数中，样本数据为给定值，而模型参数是变量。也就是说，似然分布本质上是模型参数的函数。



上一章，我们后来用二项分布作为似然分布。本章用多项分布作似然分布。二项分布可以视作是多项分布的特例。

n 为抓取动物的总数，随机变量 X_1, X_2, X_3 代表其中鸡、兔、猪数量， x_1, x_2, x_3 代表 X_1, X_2, X_3 的取值。因此，如下等式成立：

$$x_1 + x_2 + x_3 = n \quad (7)$$

在 $\theta = \theta$ 的条件下， (X_1, X_2, X_3) 满足如下多项分布：

$$f_{\chi|\Theta}(\mathbf{x}|\boldsymbol{\theta}) = f_{X_1, X_2, X_3|\Theta}(x_1, x_2, x_3|\boldsymbol{\theta}) = \frac{n!}{(x_1!) \times (x_2!) \times (x_3!)} \times \theta_1^{x_1} \times \theta_2^{x_2} \times \theta_3^{x_3} \quad (8)$$

χ 代表 X_1, X_2, X_3 构成的向量。

最大似然 MLE

似然函数 $f_{\chi|\Theta}(\mathbf{x}|\boldsymbol{\theta})$ 取对数，并忽略系数：

$$x_1 \ln \theta_1 + x_2 \ln \theta_2 + x_3 \ln \theta_3 \quad (9)$$

$\theta_1, \theta_2, \theta_3$ 存在 $\theta_1 + \theta_2 + \theta_3 = 1$ 等式约束。用拉格朗日乘子法，我们可以很容易把含约束优化问题转化为无约束问题，求得 MLE 的解为：

$$\hat{\theta}_1 = \frac{x_1}{n}, \quad \hat{\theta}_2 = \frac{x_2}{n}, \quad \hat{\theta}_3 = \frac{x_3}{n} \quad (10)$$



忘记拉格朗日乘子法读者，可以回顾《矩阵力量》第 18 章。

后验分布

后验分布代表“先验 + 数据”融合后对参数的信念。

由于后验 \propto 似然 \times 先验，后验概率 $f_{\Theta|\chi}(\boldsymbol{\theta}|\mathbf{x})$ ：

$$f_{\Theta|\chi}(\boldsymbol{\theta}|\mathbf{x}) \propto f_{\chi|\Theta}(\mathbf{x}|\boldsymbol{\theta}) f_{\Theta}(\boldsymbol{\theta}) \quad (11)$$

所以：

$$\begin{aligned} f_{\Theta|\chi}(\boldsymbol{\theta}|\mathbf{x}) &\propto \theta_1^{x_1} \times \theta_2^{x_2} \times \theta_3^{x_3} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \theta_3^{\alpha_3-1} \\ &= \theta_1^{x_1+\alpha_1-1} \times \theta_2^{x_2+\alpha_2-1} \times \theta_3^{x_3+\alpha_3-1} \end{aligned} \quad (12)$$

想要把 (12) 变成概率密度函数，我们需要一个归一化系数，使得 PDF 在整个定义域上积分为 1。

很明显，我们需要的就是如下 Beta 函数：

$$B(\alpha_1 + x_1, \alpha_2 + x_2, \alpha_3 + x_3) = B(\mathbf{x} + \boldsymbol{\alpha}) = \frac{\prod_{i=1}^K \Gamma(\alpha_i + x_i)}{\Gamma\left(\sum_{i=1}^K \alpha_i + x_i\right)} \quad (13)$$

由此可知后验分布 $f_{\Theta|X}(\boldsymbol{\theta}|\mathbf{x})$ 服从 $\text{Dir}(x_1 + \alpha_1, x_2 + \alpha_2, x_3 + \alpha_3)$, 可以写成 $\text{Dir}(\mathbf{x} + \boldsymbol{\alpha})$ 。

也就是说，在这个鸡兔猪贝叶斯推断问题中，如果先验概率为 $\text{Dir}(\boldsymbol{\alpha})$ ，则后验概率为 $\text{Dir}(\mathbf{x} + \boldsymbol{\alpha})$ 。

最大后验 MAP

对于 $\text{Dir}(\mathbf{x} + \boldsymbol{\alpha})$, X_i 的众数为：

$$\frac{x_i + \alpha_i - 1}{\sum_{k=1}^K (x_k + \alpha_k) - K} = \frac{x_i + \alpha_i - 1}{n + \alpha_0 - K}, \quad x_i + \alpha_i > 1 \quad (14)$$

这就是最大后验估计 MAP 的解析解位置所在。

当 $K = 3$ 时，最大后验 MAP 的位置为：

$$\frac{x_i + \alpha_i - 1}{n + \alpha_0 - 3} \quad (15)$$

特别地，当 $\alpha_1 = \alpha_2 = \alpha_3 = 1$ 时，最大后验 MAP 的位置为：

$$\frac{x_i}{n} \quad (16)$$

此时，MAP 的解和 MLE 的解相同。

边缘分布

根据本书第 7 章，先验分布 $\text{Dir}(\boldsymbol{\alpha})$ 的三个边缘分布分别为：

$$\begin{aligned} &\text{Beta}(\alpha_1, \alpha_0 - \alpha_1) \\ &\text{Beta}(\alpha_2, \alpha_0 - \alpha_2) \\ &\text{Beta}(\alpha_3, \alpha_0 - \alpha_3) \end{aligned} \quad (17)$$

后验分布 $\text{Dir}(\mathbf{x} + \boldsymbol{\alpha})$ 的三个边缘分布分别为：

$$\begin{aligned} &\text{Beta}\left(x_1 + \alpha_1, \alpha_0 + n - (x_1 + \alpha_1)\right) \\ &\text{Beta}\left(x_2 + \alpha_2, \alpha_0 + n - (x_2 + \alpha_2)\right) \\ &\text{Beta}\left(x_3 + \alpha_3, \alpha_0 + n - (x_3 + \alpha_3)\right) \end{aligned} \quad (18)$$

后验分布的位置

$\text{Dir}(\mathbf{x} + \boldsymbol{\alpha})$ 的三个边缘分布各自的众数分别为：

$$\frac{x_i + \alpha_i - 1}{n + \alpha_0 - 2} \quad (19)$$

它们的期望值位置为：

$$\frac{x_i + \alpha_i}{n + \alpha_0} \quad (20)$$

可见当 n 足够大时，(20) 可以用来近似 (19)。而 (19) 则可以用来近似 (14)，后验分布 MAP 优化解。

也就是说，我们可以用三个边缘 Beta 分布的期望（均值）来近似后验分布 $\text{Dir}(\mathbf{x} + \boldsymbol{\alpha})$ 的 MAP 优化解。特别是在下一章中，大家会看到我们直接用后验边缘 Beta 分布的均值作为 MAP 的优化解。

表 1 比较先验、后验分布的众数和期望。

表 1. 比较先验、后验分布的众数和期望

分布	类型	统计量	位置
$\text{Dir}(\boldsymbol{\alpha})$	先验	众数 (联合 PDF 曲面最大值)	$\frac{\alpha_i - 1}{\alpha_0 - K}, \quad \alpha_i > 1$
		期望 (联合 PDF 质心)	$\frac{\alpha_i}{\alpha_0}$
$\text{Beta}(\alpha_i, \alpha_0 - \alpha_i)$	先验边缘	众数 (先验边缘分布 PDF 曲线最大值)	$\frac{\alpha_i - 1}{\alpha_0 - 2}, \quad \alpha_i > 1$
		期望 (先验边缘分布均值)	$\frac{\alpha_i}{\alpha_0}$
$\text{Dir}(\mathbf{x} + \boldsymbol{\alpha})$	后验	众数 (联合 PDF 曲面最大值) * MAP 优化解	$\frac{x_i + \alpha_i - 1}{n + \alpha_0 - K}, \quad x_i + \alpha_i > 1$
		期望 (联合 PDF 质心) * 最大化期望值	$\frac{x_i + \alpha_i}{n + \alpha_0}$
$\text{Beta}(\alpha_i + x_i, \alpha_0 - (\alpha_i + x_i))$	后验边缘	众数 (边缘 PDF 曲线最大值)	$\frac{x_i + \alpha_i - 1}{n + \alpha_0 - 2}, \quad x_i + \alpha_i > 1$
		期望 (边缘 PDF 均值) * 常用来近似 MAP 优化解	$\frac{x_i + \alpha_i}{n + \alpha_0}$

比较 Beta 分布的众数、中位数、均值

本节最后比较 $\text{Beta}(\alpha, \beta)$ 众数、中位数、均值。

众数、中位数、均值都可以用来表征 $\text{Beta}(\alpha, \beta)$ 分布的具体位置。实际上，在贝叶斯推断中，对模型参数有三种不同的**点估计** (point estimate): 1) 后验众数，2) 后验中位数，3) 后验均值。

图 2 所示为不同 $\text{Beta}(\alpha, \beta)$ 分布众数 (蓝色划线)、中位数 (黑色划线)、均值 (红色划线)。



为了更好地理解这幅图，请大家回顾本书第 2 章介绍的有关左偏、右偏的内容。

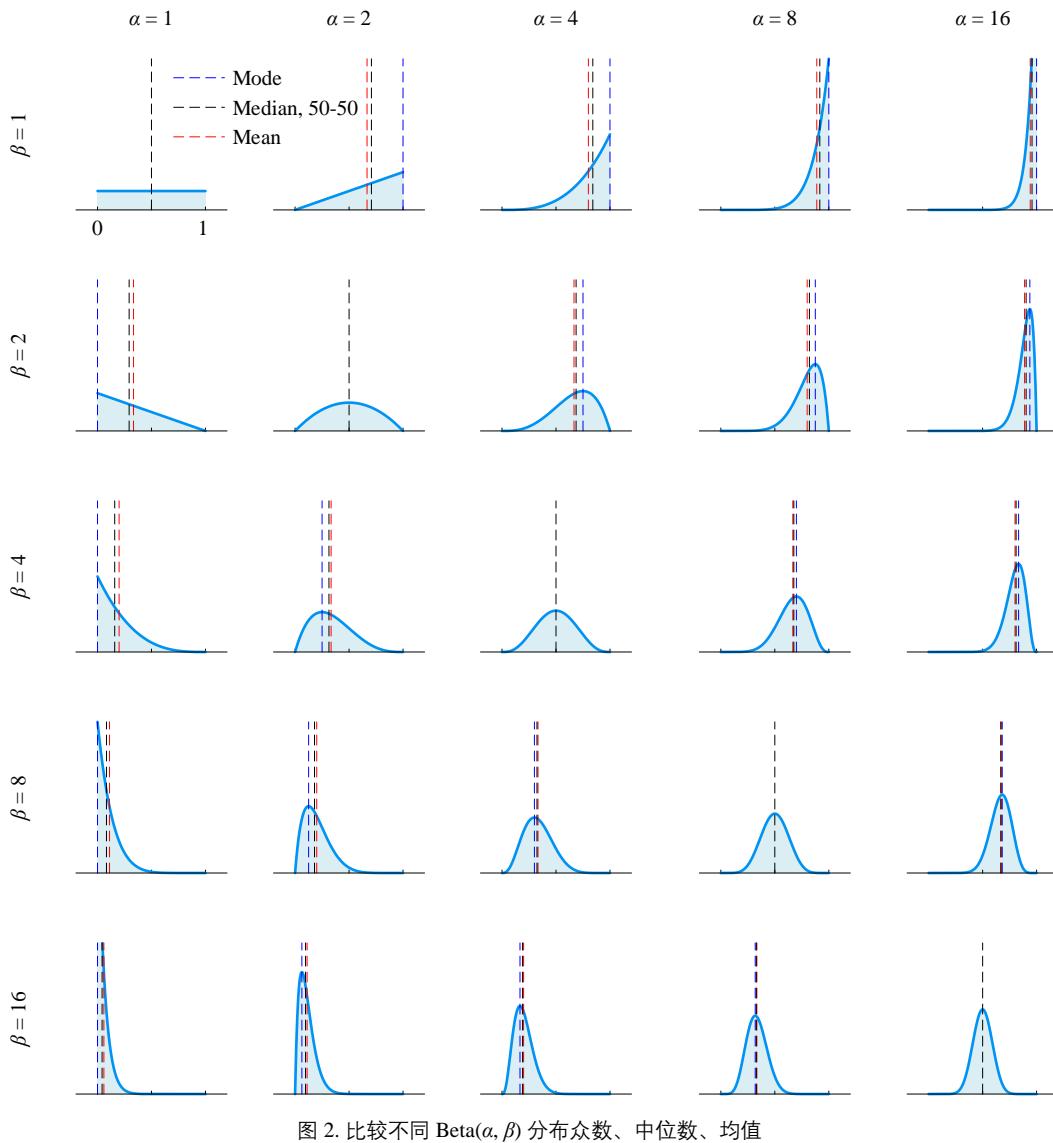
$\text{Beta}(\alpha, \beta)$ 分布的众数有明显的缺点。我们在本书第 7 章介绍过，当 α 或 β 小于等于 1 时， $\text{Beta}(\alpha, \beta)$ 的众数可能位于分布的某一端，0 或 1。比如图 2 中， $\text{Beta}(2, 1)$ 的众数位于 1，而 $\text{Beta}(1, 2)$ 的众数位于 0。这两个众数显然不能合理地表征分布的具体位置。

此外，下一章中大家会看到通过数值方法得到后验分布的曲线可能有若干局部极大值，这给 MAP 求解增加了麻烦。

因此，实践中当样本足够大时，我们常用后验边缘分布均值代替后验众数作为 MAP 的结果。

此外，后验中位数也是一个不错的选择。对于厚尾的后验分布，后验中位数要好过后验均值。因为后验均值的位置会受到厚尾影响。但是，对于蒙特卡洛模拟结果，后验中位数需要排序，计算上更困难。

特别地，如果后验分布对称，众数、均值、中位数重合。

图 2. 比较不同 Beta(α, β) 分布众数、中位数、均值

有了本节理论铺垫，下面我们结合具体实例展开讲解。本章后续三节和上一章最后三节结构相似，请大家比照阅读。

21.2 走地鸡兔猪：比例完全不确定

上一章提过，如果我们事先对动物比例值一无所知的话，我们就可以采用一个“不偏不倚”的先验分布。Dir(1, 1, 1) 显然就满足本节这个要求。这种 Dirichlet 分布又叫 flat (uniform) Dirichlet distribution。

$\text{Dir}(1, 1, 1)$ 分布概率密度值为定值，它代表我们试图保持“客观”，而不是将“主观”先验经验代入贝叶斯推断中去。图 3 所示为四种三元 Dirichlet 分布的可视化方案，本章将采用第一种， $\theta_1\theta_2$ 平面直角坐标系投影。

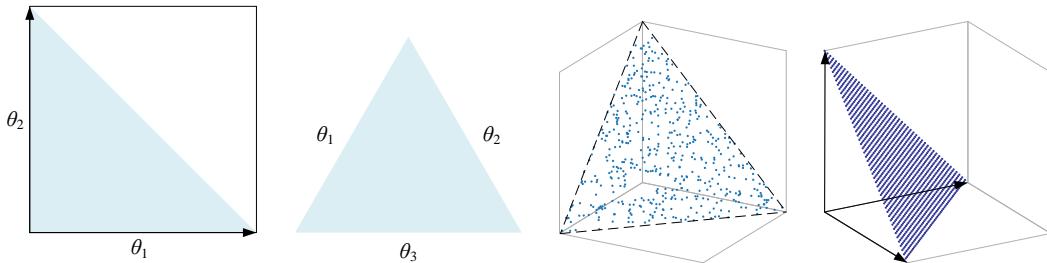


图 3. Dirichlet 分布的几种可视化方案， $\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1$

图 4 所示为某次采样的结果。图 4 (a) 中，0 代表鸡，1 代表兔，2 代表猪。

⚠ 注意，采样结果和先验分布无关。

图 4 (b) 中，随着样本数量不断增加，三种动物的比例逐渐稳定。仅仅依赖样本数据进行推断，特别是样本数量足够大时，我们已经可以得知所谓“客观”概率结果。

利用贝叶斯定理，整合“先验分布 + 样本”，我么可以得到后验分布。图 5 (a) 所示为 $\text{Dir}(1, 1, 1)$ 对应的图像。图 5 剩余 8 个不同子图展示随着样本数据 (x_1, x_2, x_3) 不断增加后验分布 $\text{Dir}(x + \alpha)$ 的变化。

图 6 所示为， n 不断增加，三个后验边缘分布位置逐渐稳定。而后验边缘分布本身变得越发“细高”，标准差不断减小，这意味着鸡兔猪的比例变得更值得信任。

图 7 比较三个不同后验边缘分布曲线形状。请大家写出每幅子图中不同后验边缘分布对应的 Beta 分布。

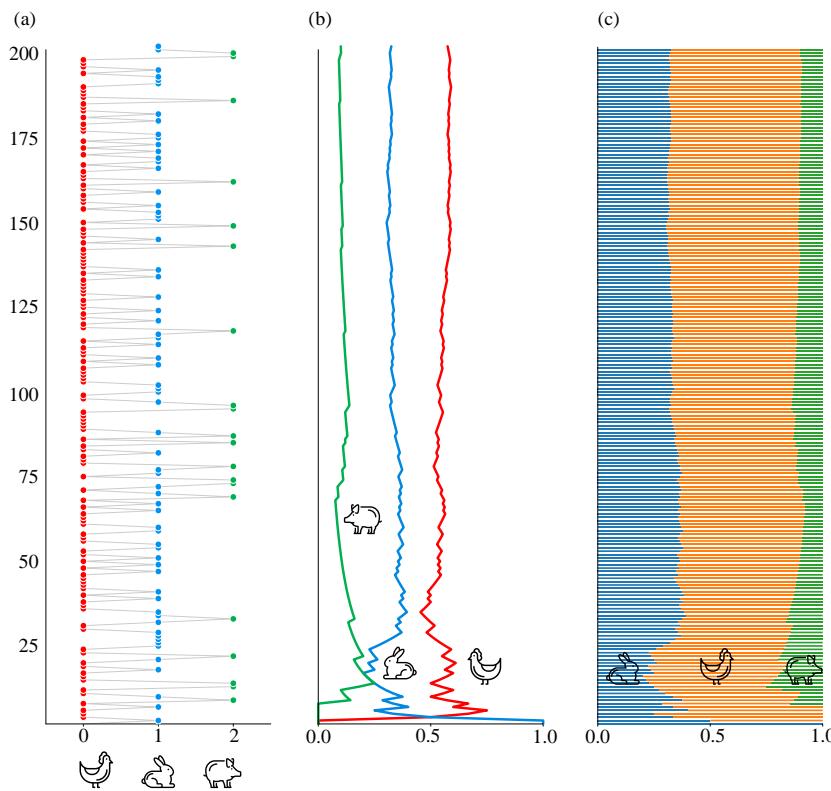
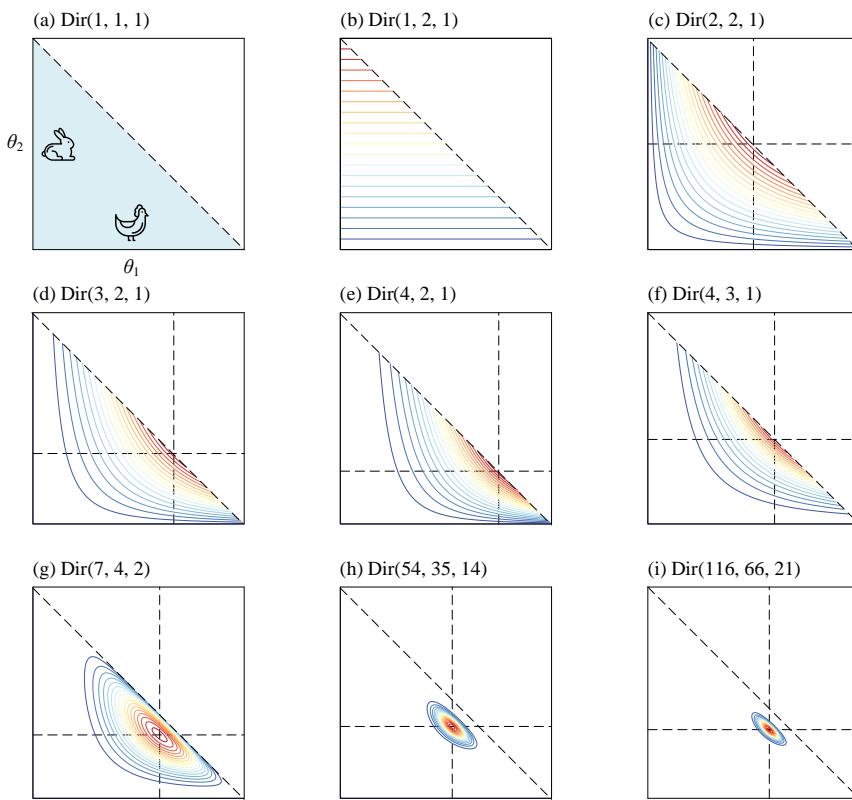


图 4. 某次试验的蒙特卡罗模拟结果



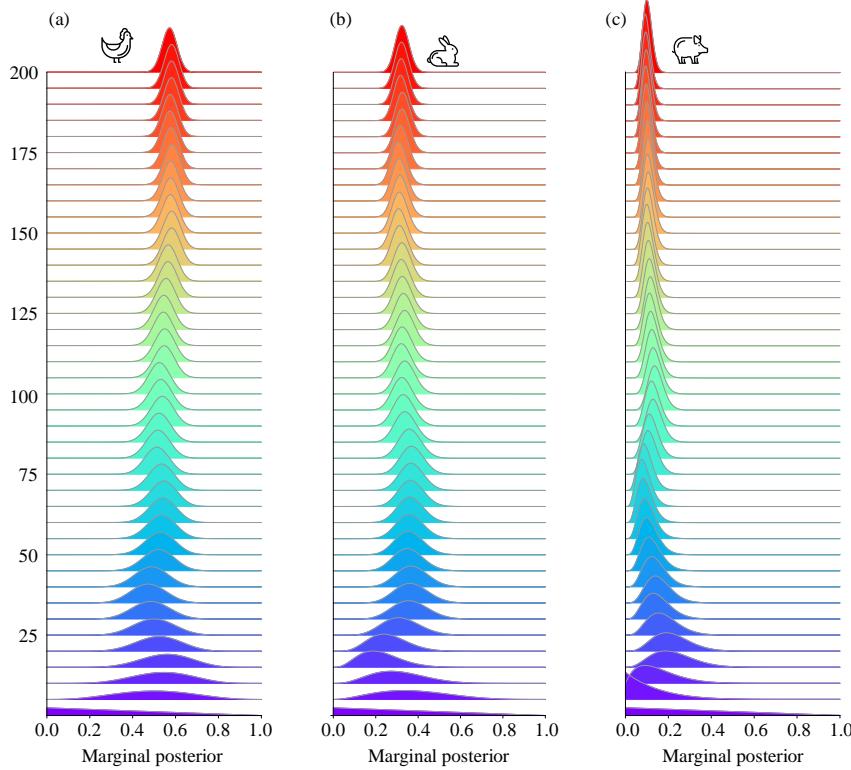
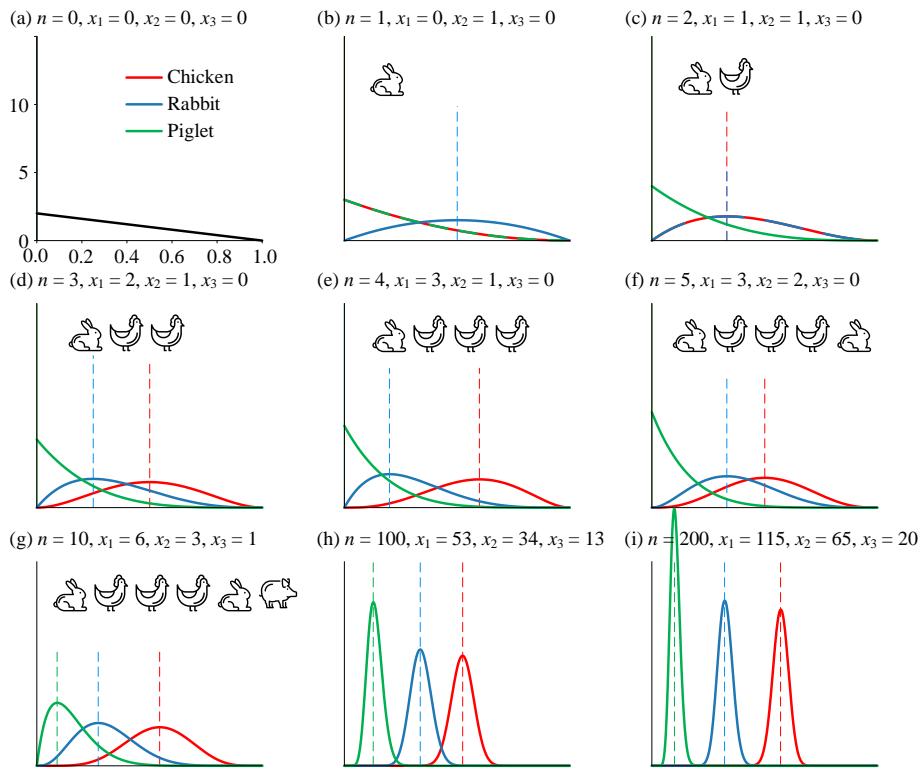
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 5. 九张 Dirichlet 分布， $\theta_1\theta_2$ 平面直角坐标系，先验分布为 $\text{Dir}(1, 1, 1)$ 图 6. 某次试验的后验边缘分布山脊图，先验分布为 $\text{Dir}(1, 1, 1)$ 图 7. 九张不同节点的后验边缘 PDF 曲线快照，先验分布为 $\text{Dir}(1, 1, 1)$

本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

21.3 走地鸡兔猪：很可能各 1/3

如果农夫认为农场的鸡兔猪的比例都是 $1/3$ ，我们就需要选用不同于前文的先验分布。这种情况，先验 Dirichlet 分布三个参数相同。

如图 8 所示为 $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 2$ 时，Dirichlet 分布的四种可视化方案。请大家分别计算 $\text{Dir}(2, 2, 2)$ 的众数、均值，并计算其边缘分布的众数、均值。

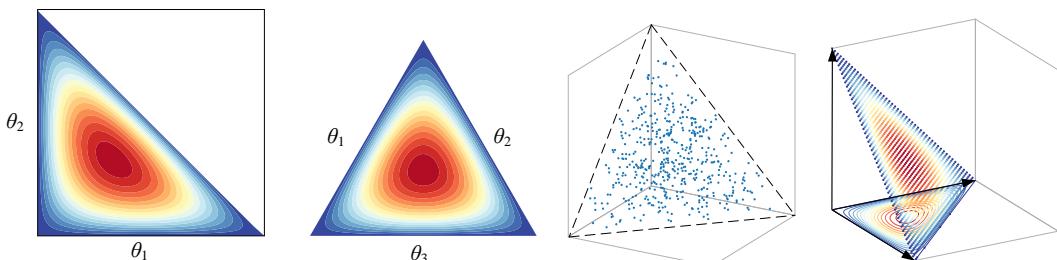


图 8. Dirichlet 分布的几种可视化方案， $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 2$

图 9 所示为 4 种不同确信度的先验分布参数设定条件下，Dirichlet 分布等高线和边缘分布曲线。图中黑色划线代表 Dirichlet 分布众数 (MAP 优化解) 所在位置。蓝色划线为边缘 Beta 分布众数位置。

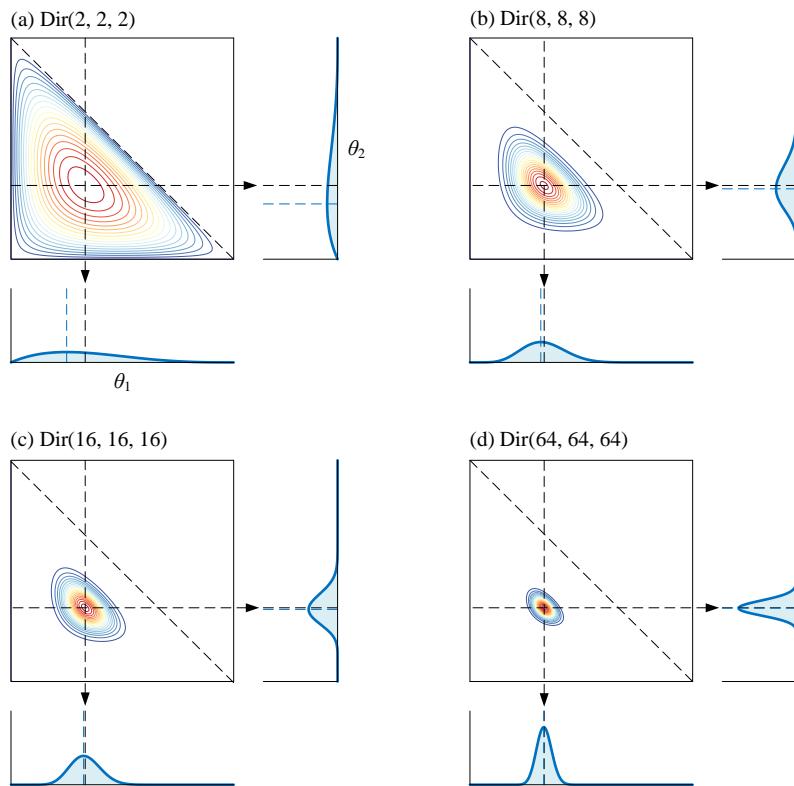


图 9. 四个不同置信度

下面，我们分两种情况完成本节蒙特卡罗模拟。随机数发生器的结果和上一节图 4 完全一致。

确信度不高

确信度不高的情况下，选择 $\text{Dir}(2, 2, 2)$ 为先验分布，如图 10 (a) 所示。

随着样本数据不断整合，图 10 剩余八幅子图所示为后验分布变化。比较图 5 (i)、图 10 (i)，可以发现样本数量较大时，后验分布受先验分布影响较小。

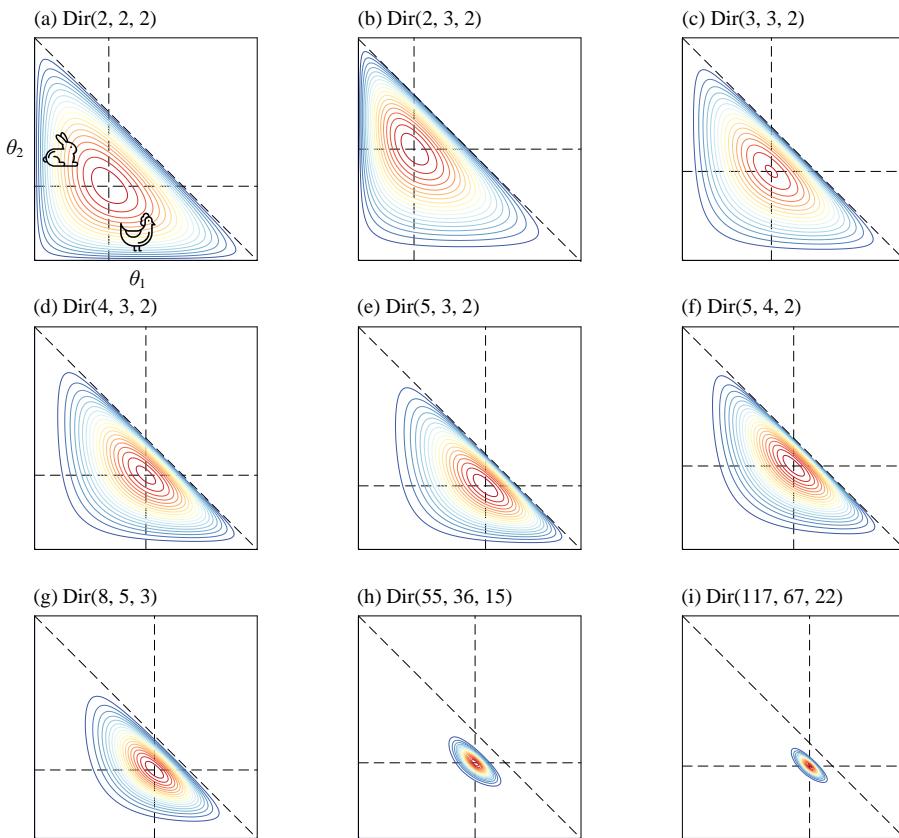
图 10. 九张 Dirichlet 分布， $\theta_1\theta_2$ 平面直角坐标系，先验分布为 $\text{Dir}(2, 2, 2)$

图 10 (g) 代表“先验 $\text{Dir}(2, 2, 2)$ + 样本 $(x_1 = 6, x_2 = 3, x_3 = 1) \rightarrow$ 后验 $\text{Dir}(8, 5, 3)$ ”。具体过程如图 11 所示。

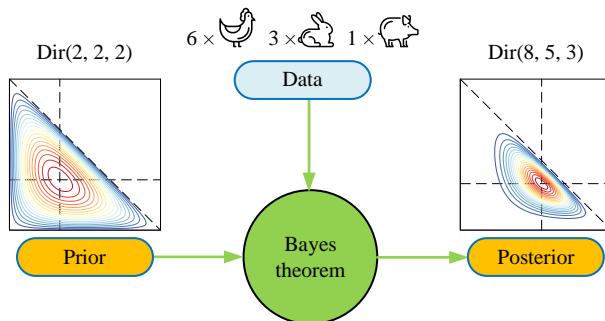
图 11. 先验 $\text{Dir}(2, 2, 2)$ + 样本 \rightarrow 后验 $\text{Dir}(8, 5, 3)$

图 12 所示为后验边缘分布的山脊图。比较图 6、图 12，容易发现当 n 比较小时，后验边缘分布曲线差异较大； n 增大后，后验边缘分布趋同。

图 13 比较三个不同的后验边缘分布。

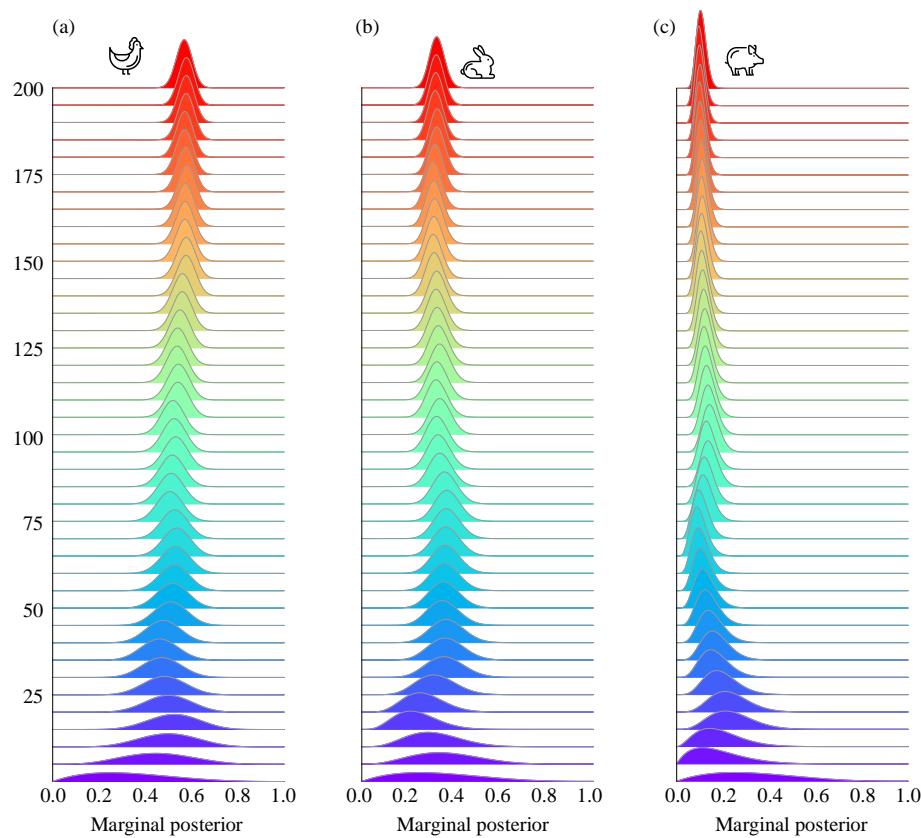
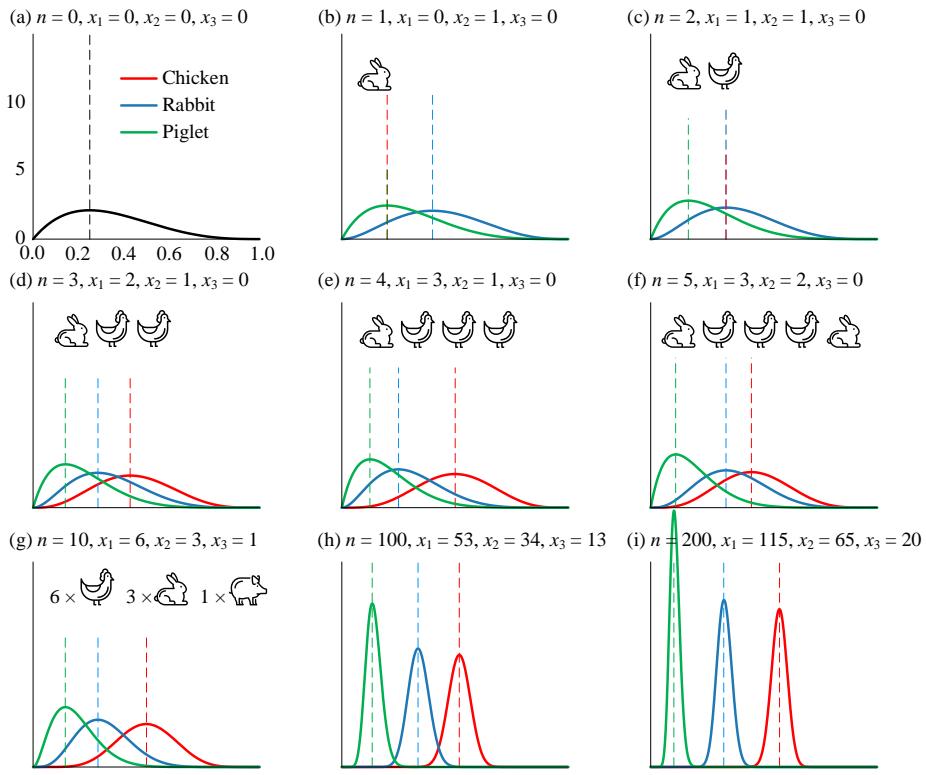


图 12. 某次试验的后验边缘分布山脊图，先验分布为 $\text{Dir}(2, 2, 2)$

图 13. 九张不同节点的后验边缘 PDF 曲线快照，先验分布为 $\text{Dir}(2, 2, 2)$

确信度很高

当农夫对 $1/3$ 的比例确信度比较高时，我们可以选择 $\text{Dir}(8, 8, 8)$ 作为先验分布。比较图 10 (a)、图 14 (a)，我们可以发现先验分布变得更加细高，这意味着边缘分布的均方差减小，确信度提高。

请大家自行分析图 14 剩余子图，并对比图 10。

图 15 先验分布为 $\text{Dir}(8, 8, 8)$ 条件下，后验边缘分布的山脊图。图 16 比较不同后验边缘分布。请大家自行分析这两图图像。

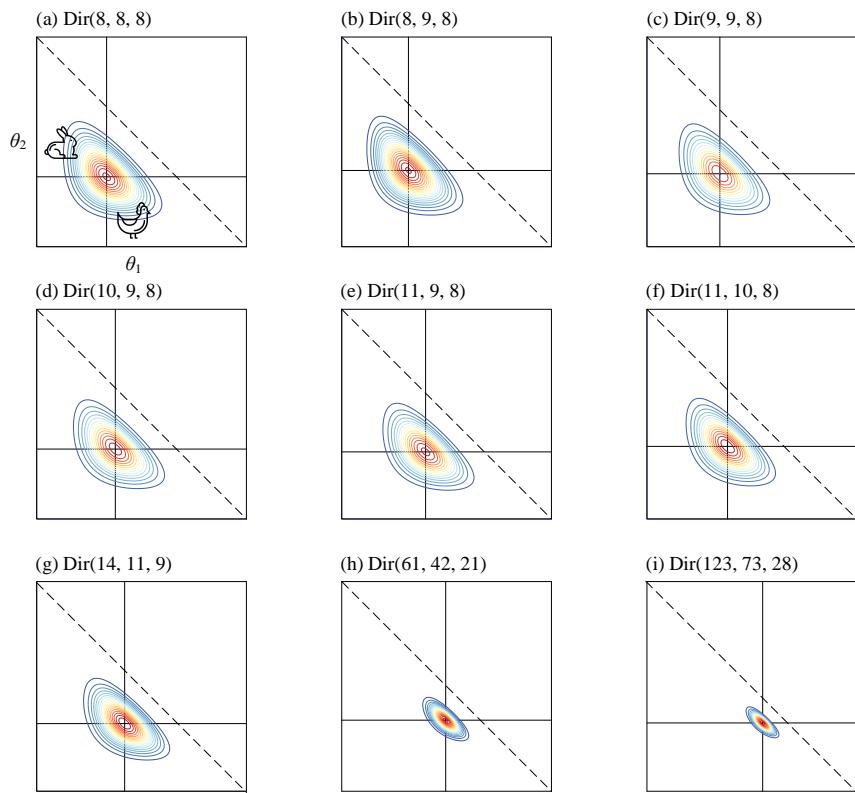
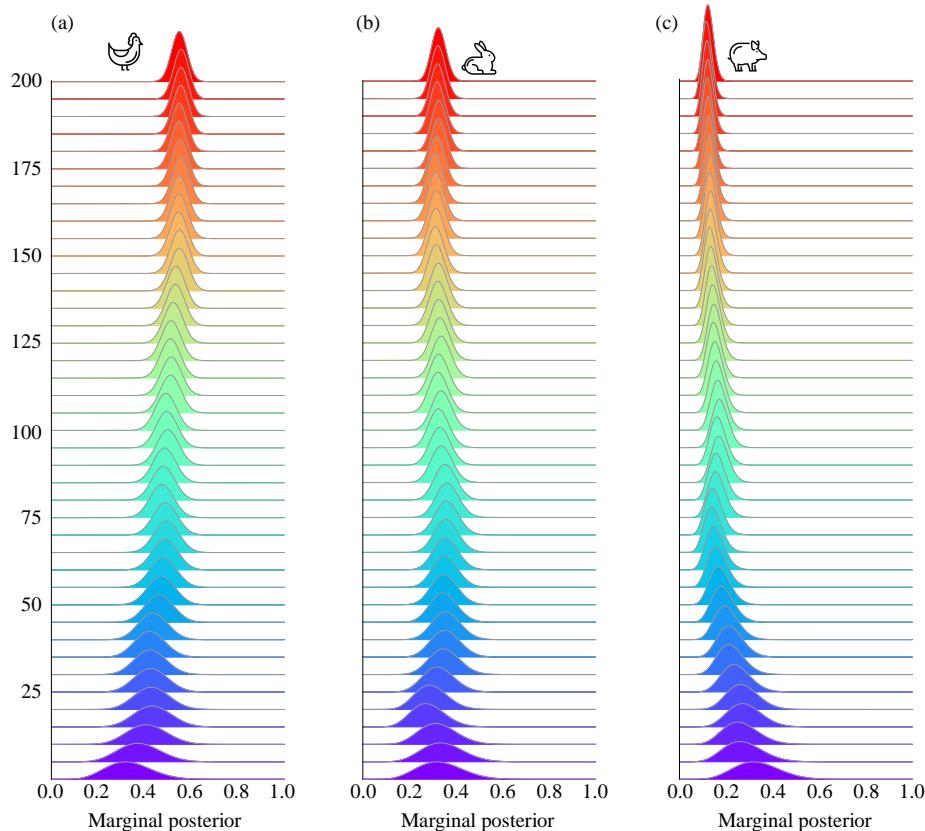


图 14. 九张 Dirichlet 分布, $\theta_1\theta_2$ 平面直角坐标系, 先验分布为 $\text{Dir}(8, 8, 8)$

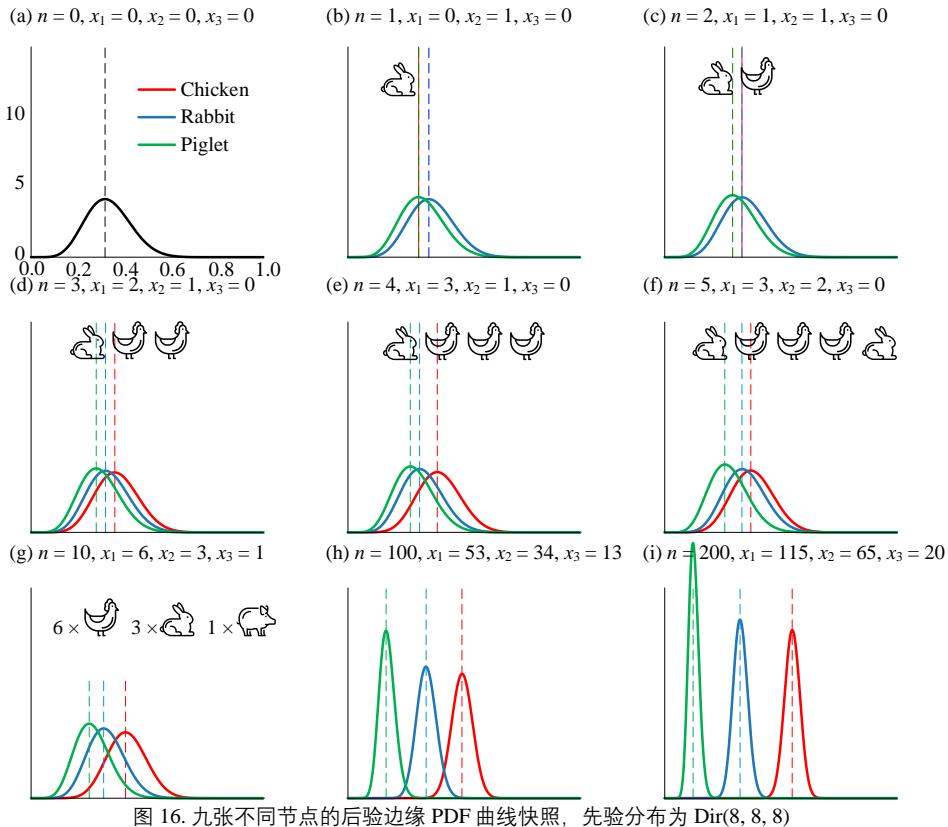
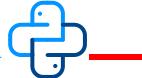


本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 15. 某次试验的后验边缘分布山脊图，先验分布为 $\text{Dir}(8, 8, 8)$ 图 16. 九张不同节点的后验边缘 PDF 曲线快照，先验分布为 $\text{Dir}(8, 8, 8)$ 

代码 Bk5_Ch21_01.py 完成本章前文蒙特卡洛模拟和可视化。

21.4 走地鸡兔猪：更一般的情况

不同先验

上一章提过，如果样本数据足够大，先验对后验的影响微乎其微。如图 17 所示，从完全不同先验出发得到的后验结果非常相似。

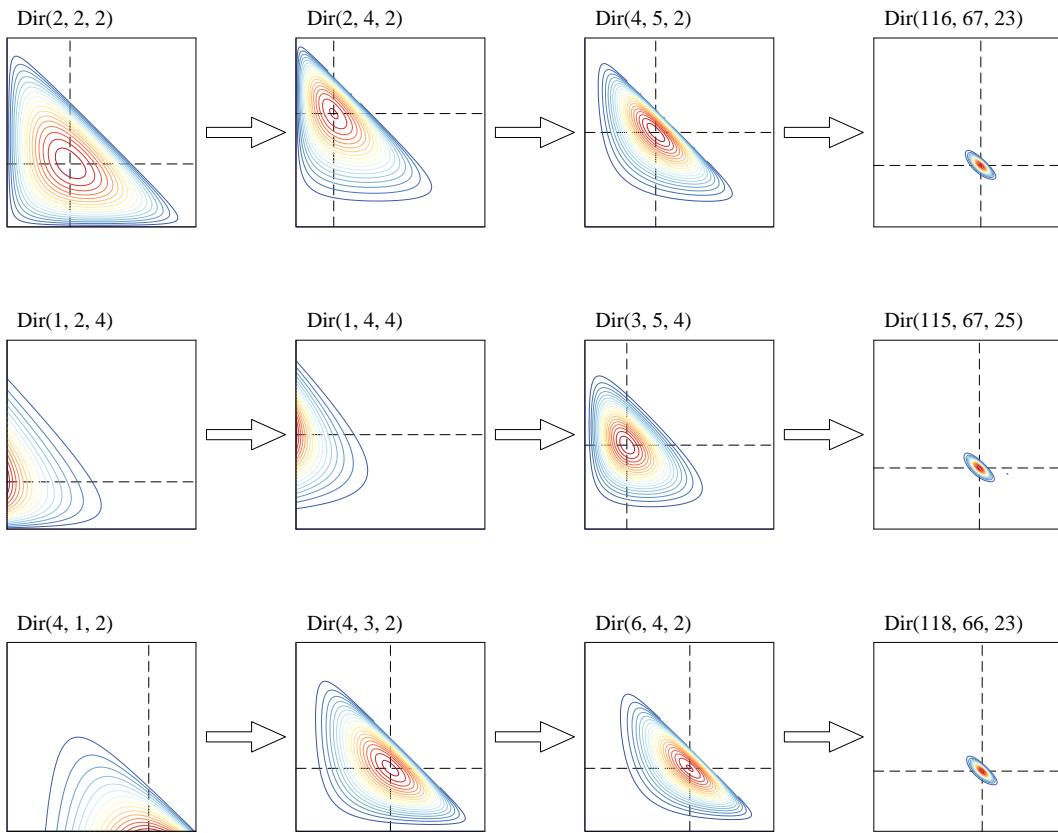


图 17. 如果样本数据足够大，先验对后验的影响微乎其微

贝叶斯收缩

上一章介绍了贝叶斯收缩，本章贝叶斯推断的结果也可以用这个视角来理解。

$\text{Dir}(\boldsymbol{x} + \boldsymbol{\alpha})$ 的后验边缘分布的期望也可以写成两部分：

$$\begin{aligned} \frac{x_i + \alpha_i}{n + \alpha_0} &= \frac{\alpha_i}{n + \alpha_0} + \frac{x_i}{n + \alpha_0} \\ &= \frac{\alpha_0}{n + \alpha_0} \times \frac{\alpha_i}{\alpha_0} + \frac{n}{n + \alpha_0} \times \frac{x_i}{n} \end{aligned} \quad (21)$$

Prior mean Sample mean

其中， $\alpha_0 = \sum_{i=1}^K \alpha_i$ ， $n = \sum_{i=1}^K x_i$ 。

以本章“鸡兔猪”为例，先验分布为 $\text{Dir}(\alpha_1, \alpha_2, \alpha_3)$ ， α_1/α_0 代表动物中鸡的比例， α_2/α_0 为兔子比例， α_3/α_0 为猪的比例。

抽取 n 只动物，其中 x_1 只鸡、 x_2 只兔、 x_3 只猪，比例分别对应 x_1/n 、 x_2/n 、 x_3/n 。

如图 18 所示，后验分布 $\text{Dir}(\alpha_1 + x_1, \alpha_2 + x_2, \alpha_3 + x_3)$ 代表“先验 + 数据”融合得到“后验”。

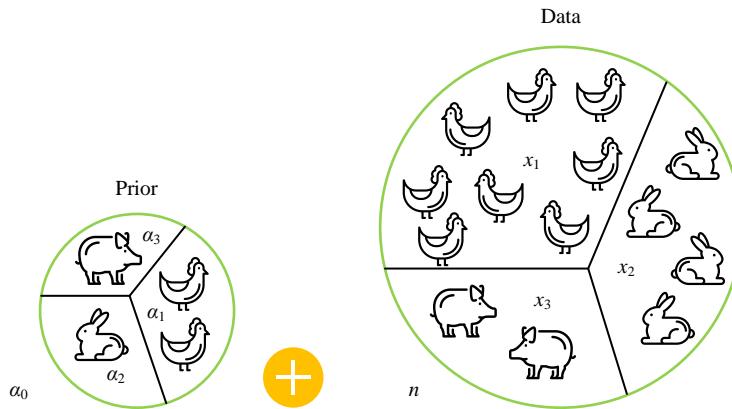


图 18. “混合”先验、样本数据

贝叶斯可信区间

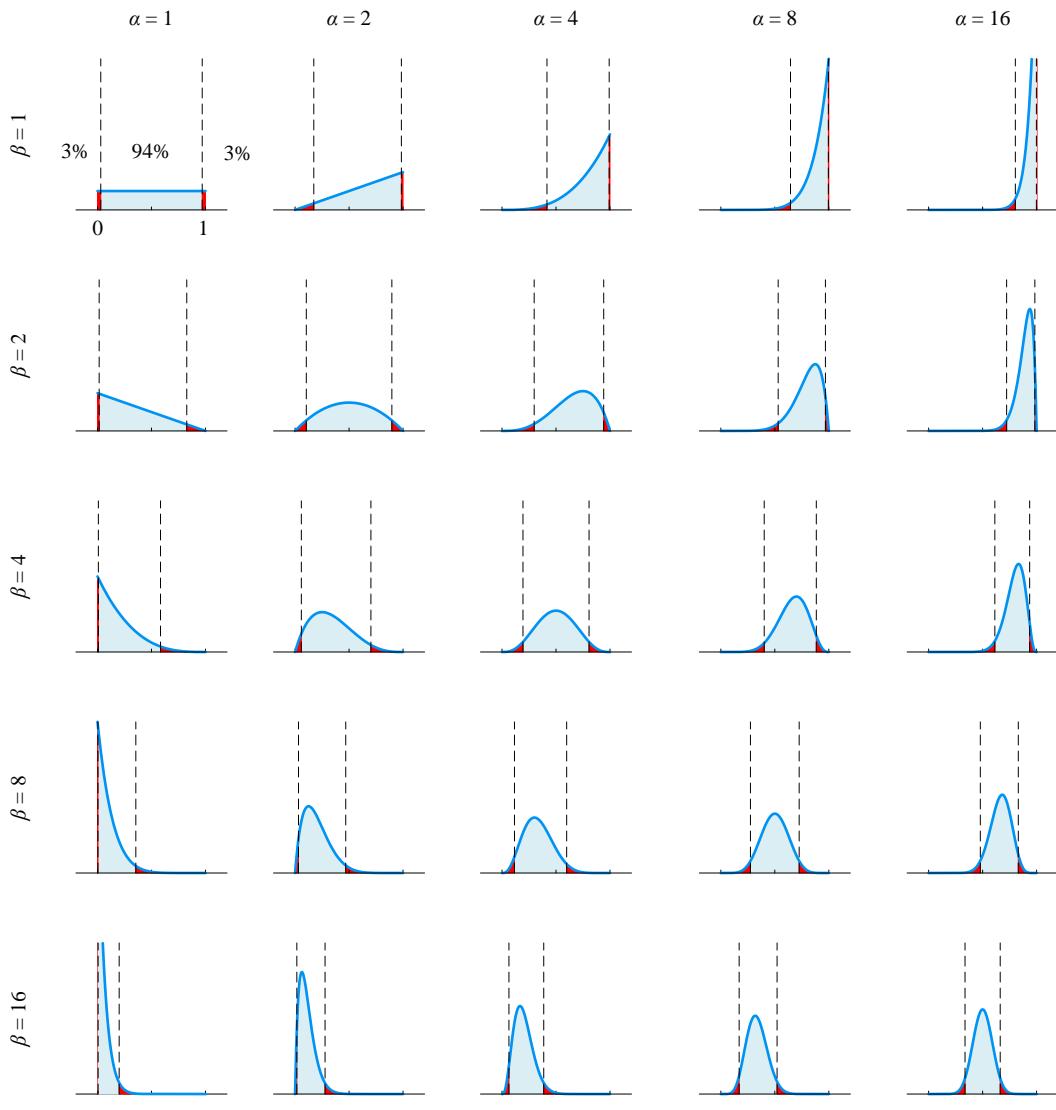
实际上，贝叶斯推断中，我们直接采用后验分布得到模型参数的各种推断，比如点估计、区间估计等等。最大化后验 MAP 就是点估计的一种。贝叶斯推断中，我们还会遇到可信区间 (credible interval)。

贝叶斯推断的可信区间不同于本书第 16 章介绍的置信区间。在频率学派中，模型参数是固定值，而样本是随机的。因此，样本的置信区间 (confidence interval) 代表参数的真实值落在该区间内的概率为 $1 - \alpha$ 。

由于贝叶斯学派认为模型参数是一个随机变量，可信区间本身就是随机变量的一个取值范围。随着样本增多，对参数信心增强，可信区间缩窄。

总结来说，置信区间是频率学派中的概念，可信区间是贝叶斯学派中的概念。置信区间是通过对样本数据进行统计分析得出的，而可信区间是通过考虑先验概率和后验概率计算得出的。置信区间是指真实参数值落在这个区间的概率，而可信区间是指这个区间的参数值有一定的可信度。置信区间的计算方法基于频率学派经典统计学理论，而可信区间的计算方法基于贝叶斯统计学理论。

下一章中，大家会发现贝叶斯推断中常用 94% 双尾可信区间。图 19 所示为不同 Beta 分布的 94% 双尾可信区间，左、右尾分别对应 3%。当概率密度曲线非对称时，我们可以发现区间左右端点对应的概率密度值一般不同。

图 19. 比较 Beta(α, β) 分布 94% 双尾可信区间

共轭先验

选择先验是有技巧的！

为了方便运算，在 $f_{\Theta|x}(\theta|x) = \frac{f_{x|\Theta}(x|\theta)f_\Theta(\theta)}{\int_\theta f_{x|\Theta}(x|\theta)f_\Theta(\theta)d\theta}$ 中，选取合适的先验分布 $f_\Theta(\theta)$ 能让后验

分布 $f_{\Theta|x}(\theta|x)$ 和先验分布 $f_\Theta(\theta)$ 具有相同的数学形式。

这就是上一章提到的，如果后验分布与先验分布属于同类，则先验分布与后验分布被称为**共轭分布** (conjugate distribution)，而先验分布被称为似然函数的**共轭先验** (conjugate prior)。

简单来说，在贝叶斯统计学中，如果我们选择先验分布和似然函数为特定的概率分布，那么我们可以计算得到一个具有相同函数形式的后验分布，这种性质被称为共轭性，对应的先验分布和后验分布就被称为共轭先验分布和共轭后验分布。

使用共轭先验，无需计算积分就可以得到后验的闭式解。我们仅仅需要跟新观察到的样本数据即可。

上一章的二项分布、Beta 分布，这一章的多项分布、Dirichlet 分布都是成对共轭分布。其他常用的成对共轭分布有：泊松分布-Gamma 分布，正态分布-正态分布，几何分布-Gamma 分布。



本章把贝叶斯推断的维度从二元提高到了三元。先验分布采用了 Dirichlet 分布，似然分布采用多项分布，而后验分布还是 Dirichlet 分布。Beta 分布可以视作 Dirichlet 分布的特例。同理，二项分布可以视作多项分布的特例。

贝叶斯推断中，后验 \propto 似然 \times 先验，无疑是最重要的关系。这个比例关系足以确定后验概率的形状，我们只需要找到一个归一化常数让后验分布在整个域上积分为 1。

本章还比较了不同 Beta 分布的众数、中位数、均值，以及它们在贝叶斯统计中的适用场合。

上一章和本章中，我们很“幸运地”避免了复杂积分运算，这是因为我们选用了共轭分布。下一章将介绍如何用马尔科夫链蒙特卡罗模拟获得后验分布。

22

Fundamentals of Markov Chain Monte Carlo

马尔科夫链蒙特卡罗

使用 PyMC3 产生满足特定后验分布的随机数



我们必须谦虚地承认，数字纯粹是人类思想的产物，但宇宙存却是颠扑不破的真理，它超然于人类思想。因此我们不能管宇宙的属性叫先验。

We must admit with humility that, while number is purely a product of our minds, space has a reality outside our minds, so that we cannot completely prescribe its properties a priori.

—— 卡尔·弗里德里希·高斯 (Carl Friedrich Gauss) | 德国数学家、物理学家、天文学家 | 1777 ~ 1855



- ◀ numpy.arange() 根据指定的范围以及设定的步长，生成一个等差数组
- ◀ numpy.concatenate() 将多个数组进行连接
- ◀ numpy.linalg.eig() 特征值分解
- ◀ numpy.random.uniform() 产生满足连续均匀分布的随机数
- ◀ numpy.zeros_like() 用来生成和输入矩阵形状相同的零矩阵
- ◀ pymc3.Dirichlet() 定义 Dirichlet 先验分布
- ◀ pymc3.Multinomial() 定义多项分布似然函数
- ◀ pymc3.plot_posterior() 绘制后验分布
- ◀ pymc3.sample() 产生随机数
- ◀ pymc3.traceplot() 绘制后验分布随机数轨迹图
- ◀ scipy.stats.beta() Beta 分布
- ◀ scipy.stats.beta.pdf() Beta 分布概率密度函数
- ◀ scipy.stats.binom() 二项分布
- ◀ scipy.stats.binom.pmf() 二项分布概率质量函数
- ◀ scipy.stats.binom.rvs() 二项分布随机数
- ◀ scipy.stats.dirichlet() Dirichlet 分布
- ◀ scipy.stats.dirichlet.pdf() Dirichlet 分布概率密度函数
- ◀ scipy.stats.norm.pdf() 正态分布概率分布 PDF
- ◀ scipy.stats.norm.ppf() 高斯分布百分点函数 PPF
- ◀ scipy.stats.norm.rvs() 生成正态分布分布随机数

22.1 归一化因子没有闭式解？

贝叶斯推断

回忆前两章贝叶斯推断中用到的贝叶斯定理：

$$\overbrace{f_{\Theta|X}(\theta|x)}^{\text{Posterior}} = \frac{\overbrace{f_{X|\Theta}(x|\theta)}^{\text{Likelihood}} \overbrace{f_\Theta(\theta)}^{\text{Prior}}}{\underbrace{f_X(x)}_{\text{Evidence}}} = \frac{\overbrace{f_{X|\Theta}(x|\theta)}^{\text{Likelihood}} \overbrace{f_\Theta(\theta)}^{\text{Prior}}}{\int_{\theta} \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} d\theta} \quad (1)$$

其中：

$f_{\Theta|X}(\theta|x)$ 为**后验概率** (posterior);

$f_{X|\Theta}(x|\theta)$ 为**似然概率** (likelihood);

$f_\Theta(\theta)$ 为**先验概率** (prior);

$f_X(x)$ 为**证据因子** (evidence), 起到归一化作用。

如图 1 所示，贝叶斯推断中最重要的比例关系就是，后验 \propto 似然 \times 先验：

$$\overbrace{f_{\Theta|X}(\theta|x)}^{\text{Posterior}} \propto \overbrace{f_{X|\Theta}(x|\theta)}^{\text{Likelihood}} \overbrace{f_\Theta(\theta)}^{\text{Prior}} \quad (2)$$

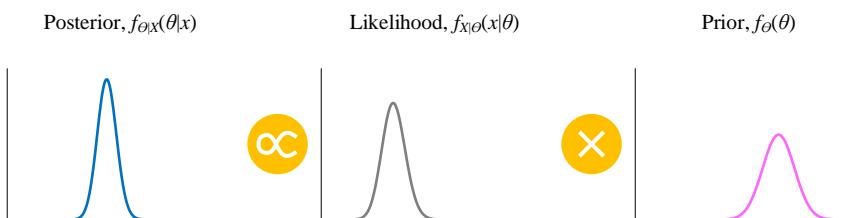
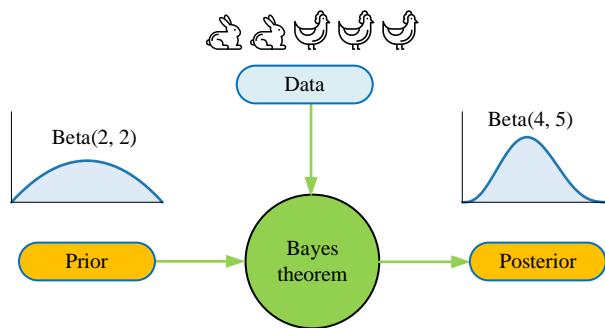


图 1. 后验 \propto 似然 \times 先验

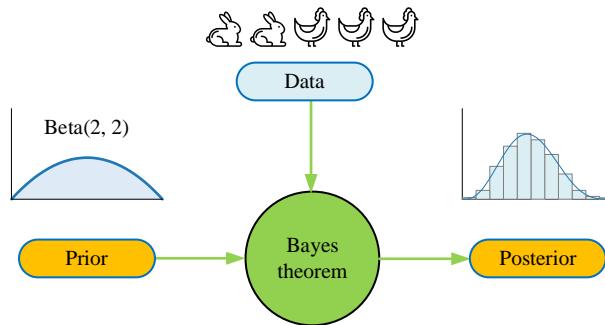
共轭分布

前两章中，如图 2 所示，我们足够“幸运”，成功地避开了 $\int_{\theta} f_{X|\Theta}(x|\theta) f_\Theta(\theta) d\theta$ 这个积分。这是因为我们选择的先验分布是似然函数的**共轭先验** (conjugate prior)，这样我们便可以得到后验概率 $f_{\Theta|X}(\theta|x)$ 的闭式解。

图 2. 先验 $\text{Beta}(2, 2)$ + 样本 (2, 3) → 后验 $\text{Beta}(4, 5)$

维数灾难

《数学要素》第 18 章介绍过数值积分。如图 3 所示，利用相同的思路，我们可以通过合理划分区间，获得后验分布的大致形状，以及对应的面积或体积，并且完成归一化。

图 3. 先验 $\text{Beta}(2, 2)$ + 样本 (2, 3) → 后验分布，数值积分

但是，这种思路仅仅适用于模型参数较小的情况。因为当模型参数很多时便会导致**维数灾难** (curse of dimensionality)。

所谓的维数灾难是指在涉及到向量的计算的问题中，随着维数的增加，计算量呈指数倍增长的一种现象。举个例子，如果模型有 3 个参数，每个参数在各自区间上均匀选取 20 个点，这个参数空间中共有 8000 个点 ($= 20 \times 20 \times 20 = 20^3$)。试想，模型如果有 20 个参数，每个维度上同样选取 20 个点，这样参数空间的点数达到惊人的 1.048×10^{26} ($= 20^{20}$)。

马尔科夫链蒙特卡洛模拟 MCMC

但是，如果我们想绕过复杂的推导过程，或者想避免数值积分带来的维数灾难，有没有其他办法获得后验分布？如图 4 所示，我们可以用**马尔科夫链蒙特卡罗模拟** (Markov Chain Monte Carlo, MCMC)。马尔科夫链蒙特卡罗模拟允许我们估计后验分布的形状，以防我们无法直接获得后验分布的闭式解。此外，蒙特卡洛方法成功地绕开了维数灾难。

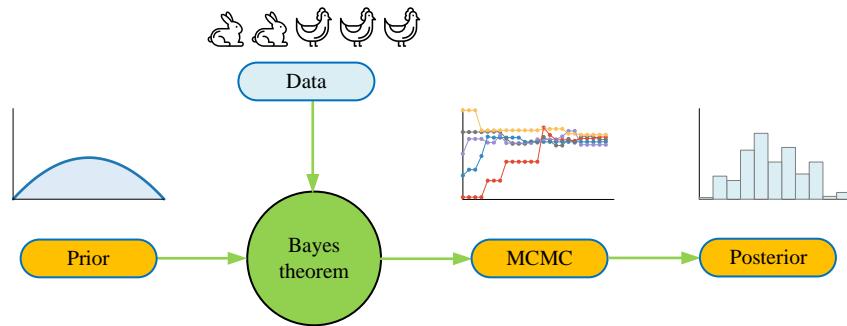


图 4. 先验 Beta(2, 2) + 样本 (2, 3) → 后验分布，马尔科夫链蒙特卡罗模拟

→ 相信大家已经发现马尔科夫链蒙特卡罗模拟有两部分——马尔科夫链、蒙特卡罗模拟。本书第 15 章专门介绍过蒙特卡洛模拟，大家对此应该很熟悉。本系列丛书的读者对“马尔科夫”这个词应该不陌生，我们在《数学要素》第 25 章“鸡兔互变”的例子中介绍过“马尔科夫”。

马尔可夫链 (Markov chain) 因俄国数学家安德烈·马尔可夫 (Andrey Andreyevich Markov) 得名，为状态空间中经过从一个状态到另一个状态的转换的随机过程。限于篇幅，本章不展开讲解马尔科夫链。

Metropolis-Hastings 采样

梅特罗波利斯-黑斯廷斯算法 (Metropolis-Hastings algorithm, MH) 是马尔可夫链蒙特卡洛中一种基本的抽样方法。

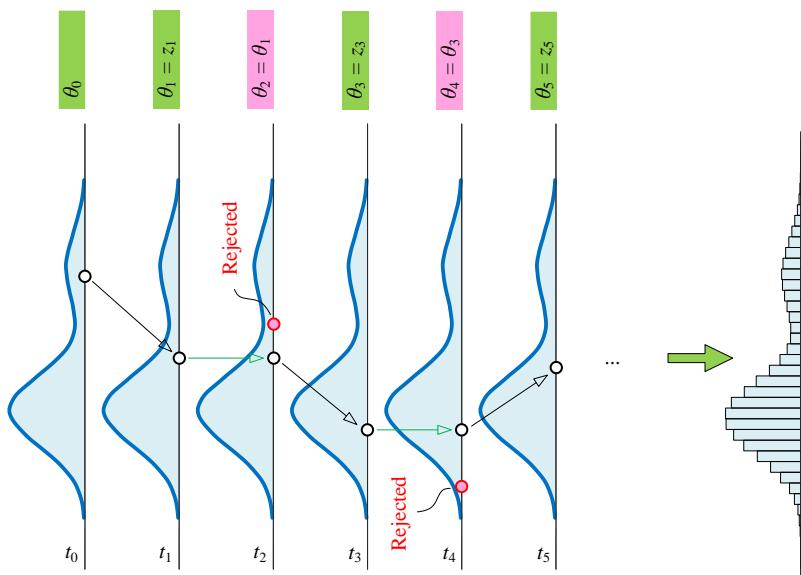


图 5. Metropolis-Hastings 采样算法原理

它通过在取值空间取任意值作为起始点，按照先验分布计算概率密度，计算起始点的概率密度。然后随机移动到下一点时，计算当前点的概率密度。移动的步伐一般从正态分布中抽取。

接着，计算当前点和起始点概率密度的比值 ρ ，并产生 $(0,1)$ 之间服从连续均匀的随机数 u 。最后，对比 ρ 与产生的随机数 u 的大小来判断是否保留当前点。当前者大于后者，接受当前点，反之则拒绝当前点。这个过程一直循环，直到获得能被接受后验分布。这一步和本书第 15 章介绍的“接受-拒绝抽样”本质上一致。

简单来说，MH 算法通过构造一个马尔可夫链，使得最终的样本分布收敛到目标分布。MH 算法核心思想是接受/拒绝准则，即通过比较接受新样本的概率与拒绝新样本的概率的比值，来决定是否接受新样本。

有关 MH 算法原理和具体流程，请大家参考李航老师的作品《机器学习方法》。

鸡兔比例

下面，我们利用 MH 算法模拟产生“鸡兔比例”中的后验分布。先验分布采用 $Beta(\alpha, \alpha)$ 。样本数据为 200 (n)，其中 60 (s) 只兔子。图 6 比较 α 取不同值时先验分布、后验分布的解析解、随机数分布。图中先验分布的随机数服从 Beta 分布，后验分布的随机数则由 MH 算法产生。

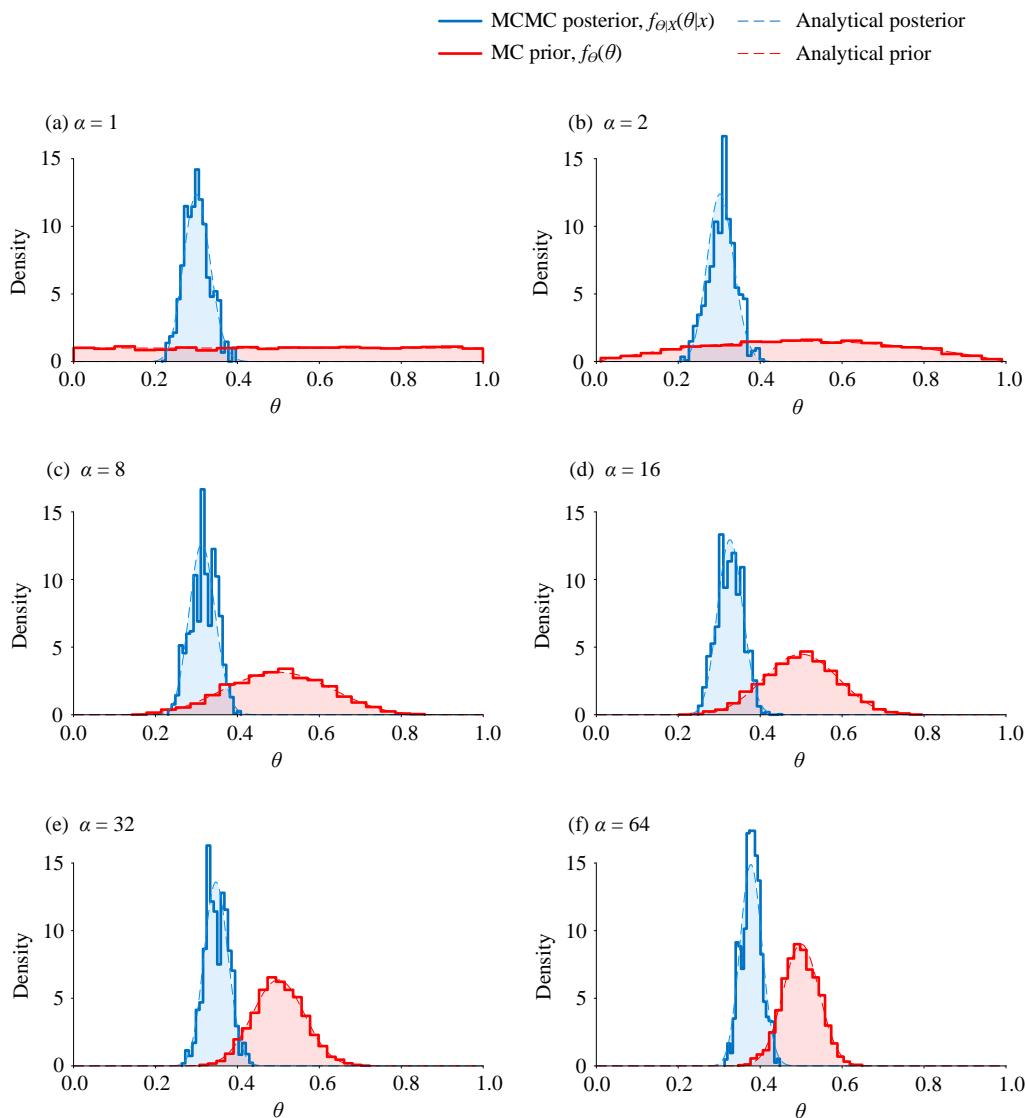
图 6. 对比先验分布、后验分布, α 取不同值时

图 7 所示为马尔科夫链蒙特卡洛模拟的收敛性。图中五条不同的后验分布随机数轨迹路径的初始值完全不同，但是它们对重都收敛于一个稳态分布，这个稳态分布对应我们要求解的后验分布。大家查看本节和本章后文代码时会发现，收敛于稳态分布之前的随机数一般都会被截断去除。

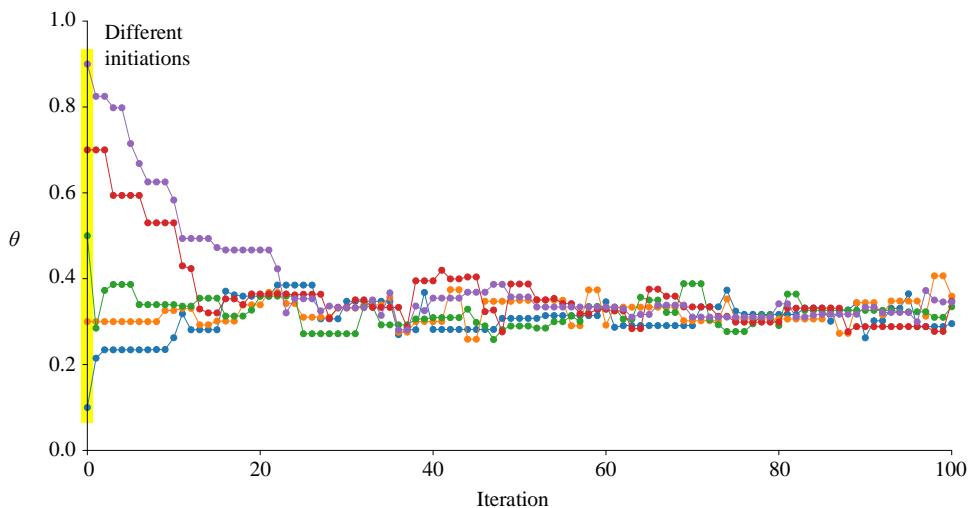


图 7. 马尔可夫链蒙特卡洛的收敛



代码 Bk5_Ch022_01.py 绘制图 6、图 7。

22.2 鸡兔比例：使用 PyMC3

本节和下一节利用 PyMC3 完成贝叶斯推断中的马尔科夫链蒙特卡罗模拟。

PyMC3 是一种 Python 开源的概率编程库，用于进行概率建模、贝叶斯统计推断和蒙特卡罗马尔可夫链蒙特卡罗 MCMC 采样。PyMC3 允许用户使用 Python 语言定义概率模型，并指定其参数的先验分布；PyMC3 支持多种先验分布，包括连续和离散分布。

PyMC3 支持使用多种 MCMC 算法进行采样，包括 NUTS、Metropolis-Hastings 和 Slice 等。PyMC3 具有丰富的可视化和后处理工具，包括 traceplot、summary、forestplot 等，方便用户对模型进行分析和诊断。

PyMC3 可以用于许多应用领域，包括机器学习、计量经济学、社会科学、物理学、生物学、神经科学等。由于 PyMC3 的简洁易用和高效性，它已经成为了许多学术界和工业界研究者进行概率建模和贝叶斯推断的首选工具之一。

先验 Beta(2,2) + 样本 2 兔 3 鸡

如图 8 所示，根据本书第 20 章内容，对于鸡兔比例问题，我们知道当先验分布为 $\text{Beta}(2, 2)$ ，引入样本数据（2 兔、3 鸡），得到的后验分布为 $\text{Beta}(4, 5)$ 。先验分布 $\text{Beta}(2, 2)$ 的均值、众数都位于 $1/2$ ，也就是鸡兔各占 50%，但是确信度不高。请大家自己计算 $\text{Beta}(4, 5)$ 均值位置。

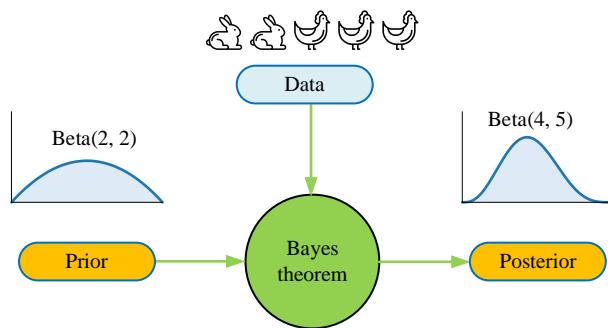


图 8. 先验 $\text{Beta}(2, 2) + \text{样本 } (2, 3) \rightarrow \text{后验 } \text{Beta}(4, 5)$

下面，我们利用 PyMC3 模拟产生这个后验分布。由于 Beta 分布是 Dirichlet 分布的特例。本节的先验分布实际上是二元 Dirichlet 分布，所以我们会看到两个后验分布。图 9 (b) 所示为后验分布随机数轨迹图，这些随机数便构成后验分布。

轨迹图中蓝色曲线对应图 9 (a) 中蓝色后验分布，即兔子比例。轨迹图中橙色曲线对应图 9 (a) 中橙色后验分布，即鸡的比例。在代码中，大家会看到随机数轨迹实际上是由两条轨迹合并而成。

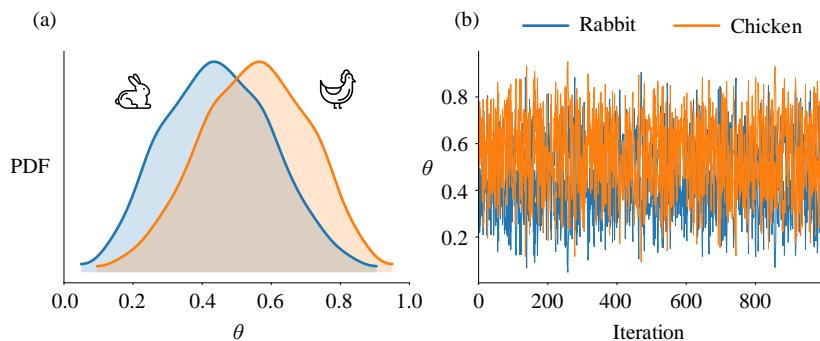


图 9. 后验分布随机数轨迹图，先验 $\text{Beta}(2,2) + \text{样本 } 2 \text{ 兔 } 3 \text{ 鸡}$

图 10 分别用直方图、KDE 曲线可视化两个后验分布。图 10 给出的均值所在位置就相当于最大后验 MAP 的优化解。

图中 HDI 代表最大密度区间 (highest density interval)。HDI 又叫 HPDI (highest posterior density interval)，本质上是上一章介绍的后验分布可信区间。

HDI 的特点是，相同置信度下，HDI 区间宽度最短，HDI 区间两端对应概率密度值相等。但是，HDI 左右尾对应的面积很可能不相等，这一点明显不同于可信区间。

图 10 (a) 告诉我们兔比例的后验分布 94% 最大密度区间的宽度为 $0.57 (= 0.75 - 0.18)$ 。鸡比例的后验分布 94% 最大密度区间的宽度也是 $0.57 (= 0.82 - 0.25)$ 。这个宽度可以用来度量确信程度。

再次强调，贝叶斯派认为模型参数本身不确定，也服从某种分布。因此可信区间或 HDI 本身就是模型参数的分布。这一点完全不同于频率派的置信区间。

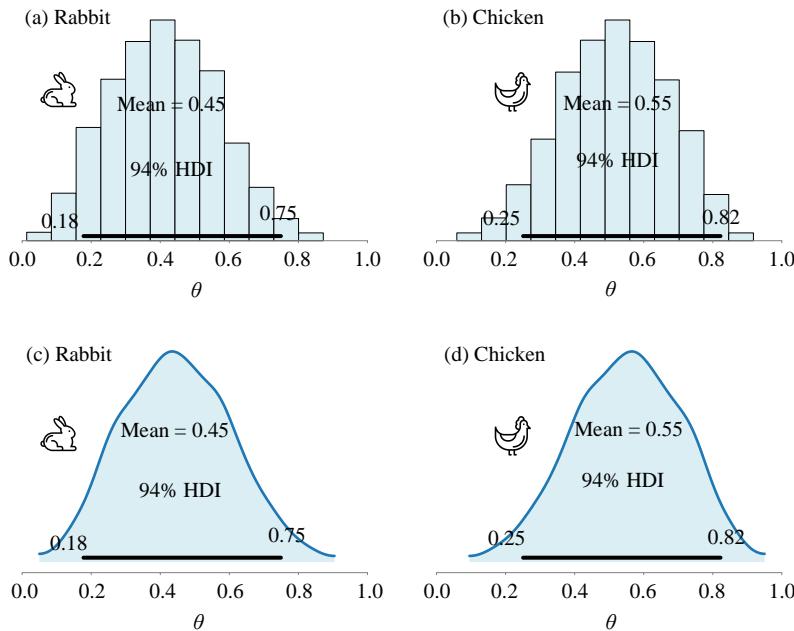


图 10. 后验分布直方图、KDE, 先验 Beta(2,2) + 样本 2 兔 3 鸡

先验 Beta(2,2) + 样本 90 兔 110 鸡

再看一个例子。如图 11 所示，先验分布还是 Beta(2, 2)，但是样本数据为 90 只兔、110 只鸡。请大家试着自己推到得到后验分布的解析式。

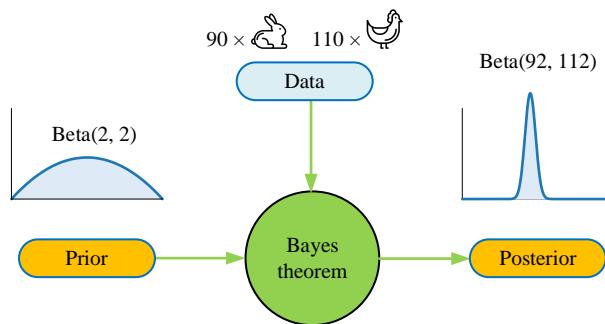


图 11. 先验 Beta(2, 2) + 样本 (90, 110) → 后验 Beta(92, 112)

图 12 (a) 所示为鸡兔比例的后验分布。图 12 (b) 所示为产生后验分布的随机数。

图 13 所示为后验分布的直方图和 KDE 曲线。虽然先验分布相同，由于引入更多样本，相比图 10，图 13 的后验分布变得更加“细高”，也就是说确信度变得更强。

图 13 (a) 告诉我们兔比例的后验分布 94% HDI 的宽度为 0.13 (= 0.51 – 0.38)。鸡比例的后验分布 94% HDI 的宽度也是 0.13 (= 0.62 – 0.49)。相比图 10，最大密度区间宽度明显缩小。

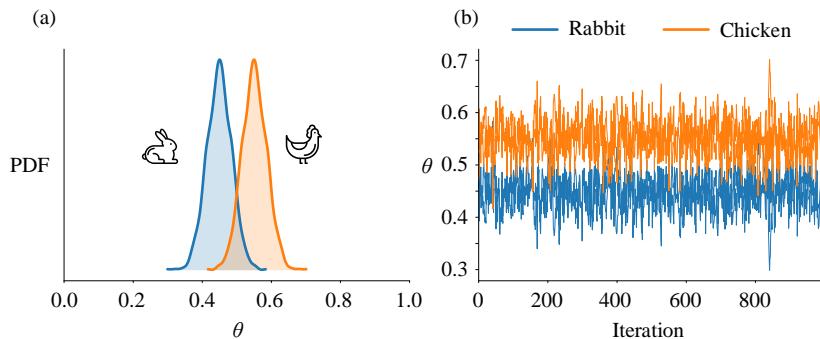


图 12. 后验分布随机数轨迹图，先验 Beta(2,2) + 样本 90 兔 110 鸡

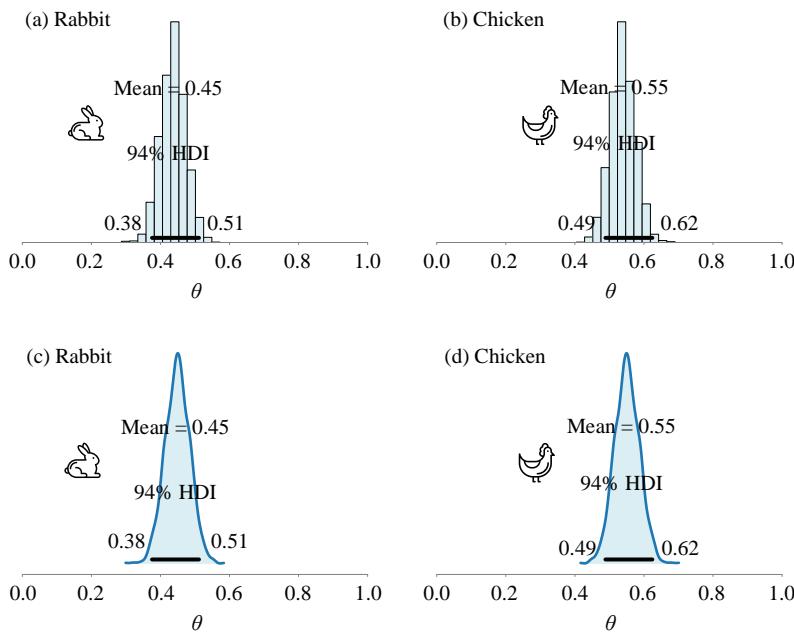
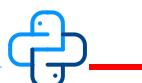


图 13. 后验分布直方图、KDE，先验 Beta(2,2) + 样本 90 兔 110 鸡



代码 Bk5_Ch22_02.ipynb 绘制图 9、图 10、图 11、图 12。请大家用 JupyterLab 打开并运行代码文件。此外，请大家改变先验分布的参数设置，并观察后验分布的变化。

22.3 鸡兔猪比例：使用 PyMC3

本节用 PyMC3 求解鸡兔猪比例的贝叶斯推断问题。

先验 $\text{Dir}(2, 2, 2)$ + 样本 3 兔 6 鸡 1 猪

选取 $\text{Dir}(2, 2, 2)$ 作为先验分布，这意味着事先主观经验认为鸡兔猪的占比都是 $1/3$ ，但是确信度不够强。如图 14 所示，观察到的 10 只动物中有 6 只鸡、3 只兔、1 只猪。利用上一章内容，我们可以推导得到后验分布为 $\text{Dir}(8, 5, 3)$ 。下面，这一节也用 PyMC3 完成 MCMC 模拟并生成后验边缘分布。

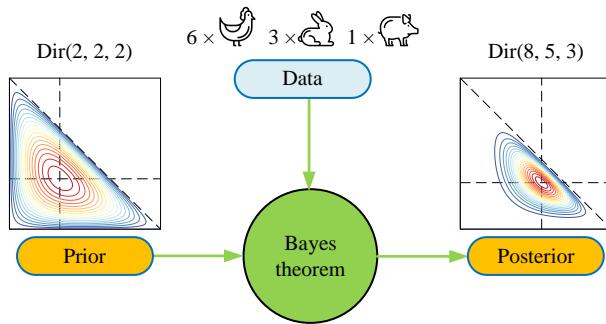


图 14. 先验 $\text{Dir}(2, 2, 2)$ + 样本 → 后验 $\text{Dir}(8, 5, 3)$

图 15 (b) 所示为后验分布随机数轨迹图，由此得到图 15 (a) 左图的后验分布。

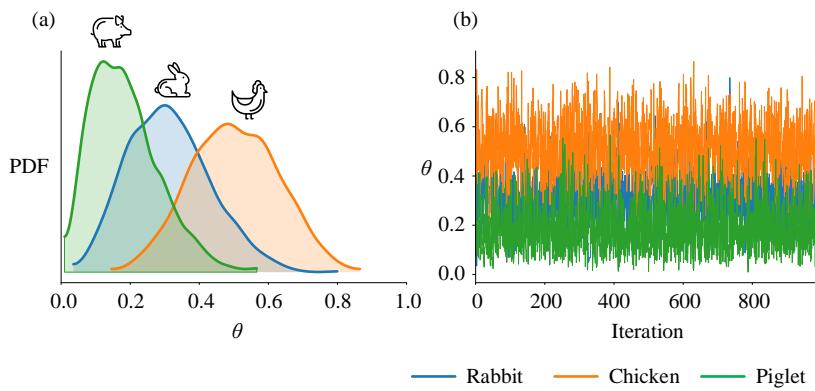
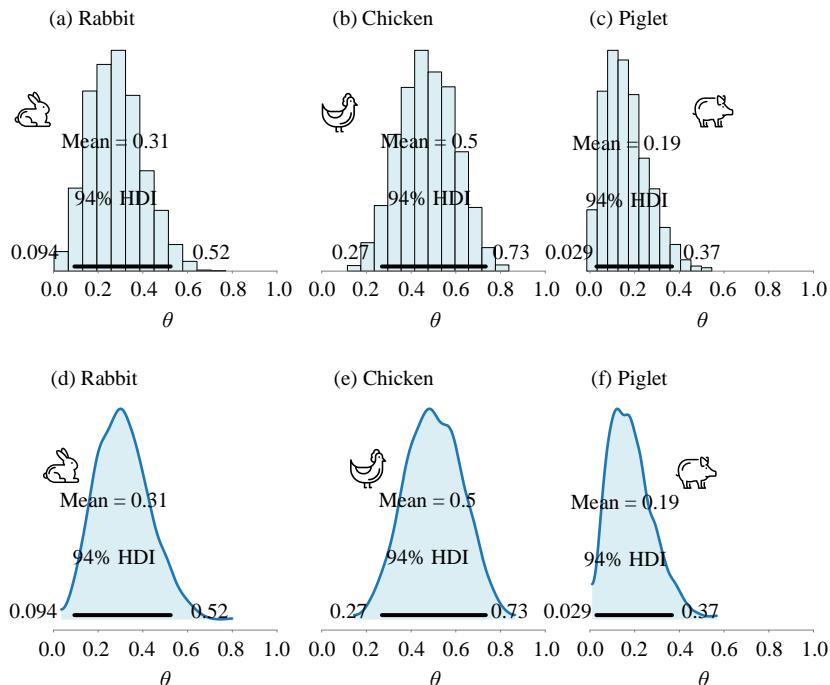


图 15. 后验分布随机数轨迹图，先验 $\text{Dir}(2,2,2)$ + 样本 3 兔 6 鸡 1 猪

图 16 所示为三种动物比例的后验分布直方图、KDE 曲线。

图 16. 后验分布直方图、KDE，先验 $\text{Dir}(2,2,2)$ + 样本 3 兔 6 鸡 1 猪

先验 $\text{Dir}(2,2,2)$ + 样本 65 兔 115 鸡 20 猪

下面保持先验分布 $\text{Dir}(2, 2, 2)$ 不变，增加样本数量（115 鸡、65 兔、20 猪），得到的后验分布为 $\text{Dir}(117, 67, 22)$ 。建议大家自己试着推导后验分布闭式解。

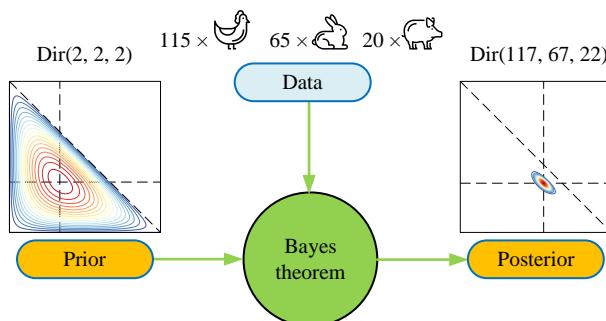
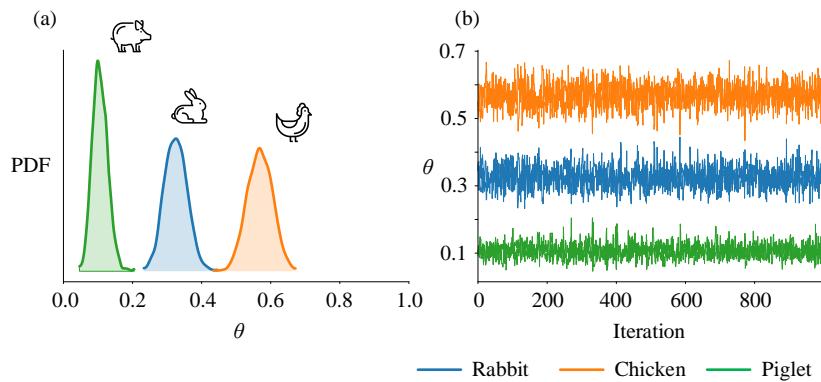
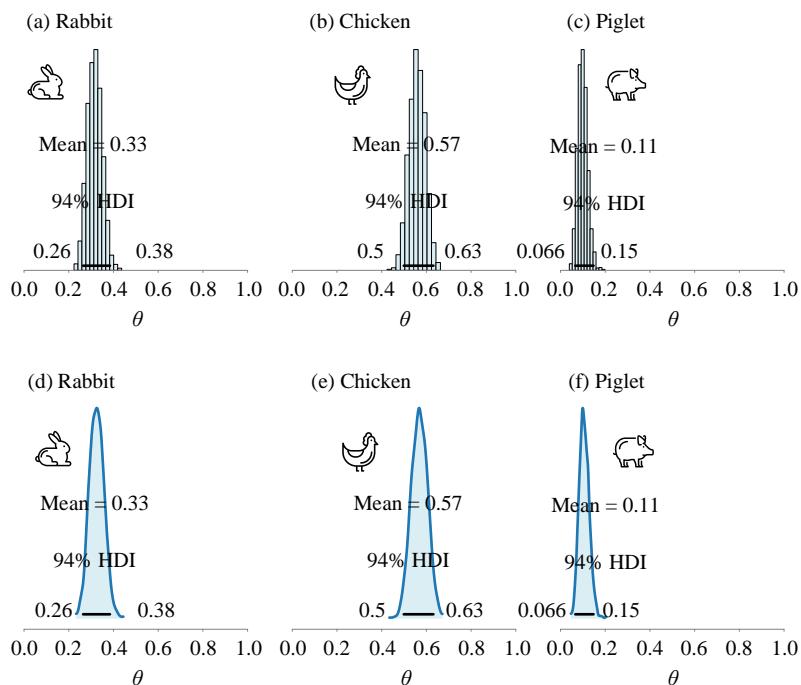
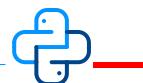
图 17. 先验 $\text{Dir}(2, 2, 2)$ + 样本 → 后验 $\text{Dir}(117, 67, 22)$

图 18 所示为三种动物的后验概率随机数的轨迹图和分布。图 19 所示为后验分布的直方图、KDE 曲线。请大家自己计算并对比图 16 和图 19 中 94% HDI 宽度。

图 18. 后验分布随机数轨迹图, 先验 $\text{Dir}(2,2,2)$ + 样本 65 兔 115 鸡 20 猪图 19. 后验分布直方图、KDE, 先验 $\text{Dir}(2,2,2)$ + 样本 65 兔 115 鸡 20 猪

代码 Bk5_Ch22_03.ipynb 绘制图 15、图 16、图 18、图 19。请大家用 JupyterLab 打开并运行代码文件。请大家改变先验分布参数，从而调整置信度，并观察后验分布的变化。

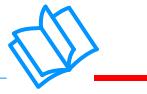


总结来说，贝叶斯推断把总体的模型参数看作随机变量。在得到样本之前，根据主观经验和既有知识给出未知参数的概率分布，称为先验分布。从总体中得到样本数据后，根据贝叶斯定理，基于给定的样本数据，得出模型参数的后验分布。并根据参数的后验分布进行统计推断。贝叶斯推断对应的优化问题为最大化后验概率，即 MAP。

在贝叶斯推断中，我们关注的核心是模型参数的后验分布。而样本数据服从怎样的分布不是贝叶斯推断关注的重点。

贝叶斯推断也并不完美！明显的缺点之一就是分析推导过程十分复杂。先验分布的建立，需要丰富的经验。采用马尔科夫链蒙特卡罗模拟，可以避免复杂推导，避免数值积分可能带来的维度灾难，但是计算成本显然较高。

读到这里，我们已经完成本书“贝叶斯”板块的学习。下面进入“椭圆三部曲”，鸢尾花书数学板块的收官之旅。



想深入学习贝叶斯推断的读者可以参考开源图书 *Bayesian Methods for Hackers: Probabilistic Programming and Bayesian Inference*:

<https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers>



Mahalanobis Distance

马氏距离

一种和椭圆有关、考虑数据分布的距离度量



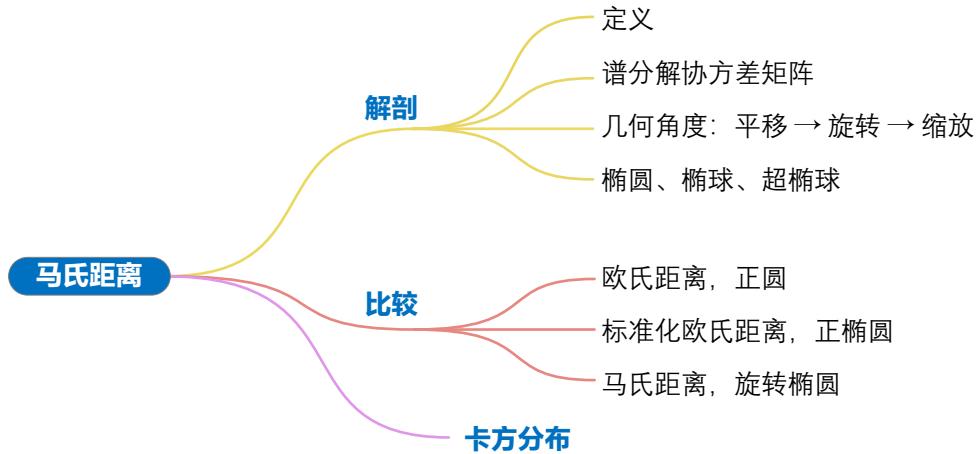
耐心，坚持！今天的苦，就是明天的甜。

Be patient and tough; someday this pain will be useful to you.

——奥维德 (Ovid) | 古罗马诗人 | 43 BC ~ 17/18 AD



- ◀ `numpy.linalg.eig()` 特征值分解
- ◀ `scipy.stats.distributions.chi2.cdf()` 卡方分布的 CDF
- ◀ `scipy.stats.distributions.chi2.ppf()` 卡方分布的百分点函数 PPF
- ◀ `seaborn.pairplot()` 成对散点图
- ◀ `seaborn.scatterplot()` 绘制散点图
- ◀ `sklearn.covariance.EmpiricalCovariance()` 估算协方差的对象，可以用来计算马氏距离



23.1 马氏距离：考虑数据分布的距离度量

本书最后三章叫做“椭圆三部曲”，我们将介绍马氏距离、线性回归、主成分分析这三个和椭圆直接有关的话题。

“鸢尾花书”的读者对马氏距离应该完全不陌生，本章将系统地讲解马氏距离及其应用。

定义

马氏距离 (Mahalanobis distance, Mahal distance)，也称**马哈距离**，具体定义如下：

$$d = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \quad (1)$$

其中， $\boldsymbol{\Sigma}$ 为样本数据 \mathbf{X} 方差协方差矩阵， $\boldsymbol{\mu}$ 为 \mathbf{X} 的质心。

注意，马氏距离的单位为标准差。

从几何来讲， d 为定值时，(1) 为质心位于 $\boldsymbol{\mu}$ 的椭圆、椭球或超椭球。

平移 → 旋转 → 缩放

对 $\boldsymbol{\Sigma}$ 谱分解得到：

$$\boldsymbol{\Sigma} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \quad (2)$$

利用 (2) 获得 $\boldsymbol{\Sigma}^{-1}$ 的特征值分解：

$$\boldsymbol{\Sigma}^{-1} = \mathbf{V} \boldsymbol{\Lambda}^{-1} \mathbf{V}^T \quad (3)$$

将 (3) 代入 (1) 整理得到：

$$d = \left\| \boldsymbol{\Lambda}^{\frac{-1}{2}} \mathbf{V}^T \begin{pmatrix} \mathbf{x} - \boldsymbol{\mu} \\ \text{Scale} \\ \text{Rotate} \\ \text{Centralize} \end{pmatrix} \right\| \quad (4)$$

其中， $\boldsymbol{\mu}$ 完成**中心化** (centralize)， \mathbf{V} 矩阵完成**旋转** (rotate)， $\boldsymbol{\Lambda}^{\frac{-1}{2}}$ 矩阵完成**缩放** (scale)。整个几何变换过程如图 1 所示。观察上式，大家已经发现马氏距离本身也是个范数。



对这部分内容感到陌生的读者，请参考本书第 11 章。大家如果忘记特征值分解、谱分解相关内容，请回顾《矩阵力量》第 13、14 章。

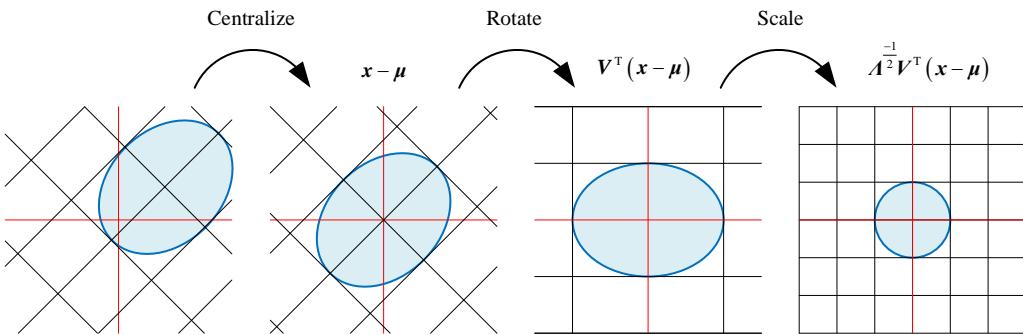


图 1. 几何变换：平移 → 旋转 → 缩放

马氏距离将协方差矩阵 Σ 纳入距离度量计算。马氏距离相当于对欧氏距离的一种修正，马氏距离完成数据**正交化** (orthogonalization)，解决特征之间相关性问题。同时，马氏距离内含**标准化** (standardization)，解决了特征之间尺度和单位不一致问题。

单特征

特别地，当特征数 $D = 1$ 时：

$$\mathbf{x} = [x], \quad \boldsymbol{\mu} = [\mu], \quad \boldsymbol{\Sigma} = [\sigma^2] \quad (5)$$

代入 (1) 得到：

$$d = \sqrt{(x - \mu) \frac{1}{\sigma^2} (x - \mu)} = \left| \frac{x - \mu}{\sigma} \right| \quad (6)$$

大家是不是觉得眼前一亮，这正是 Z 分数的绝对值， d 的单位正是标准差。如图 2 (a) 所示，比如 $d = 3$ ，意味着马氏距离为“3 个标准差”。

当特征数 $D = 2$ 时，如图 2 (b) 所示，马氏距离的几何形态是同心椭圆。当特征数 $D = 3$ 时，如图 2 (c) 所示，马氏距离的几何形态是同心椭球。

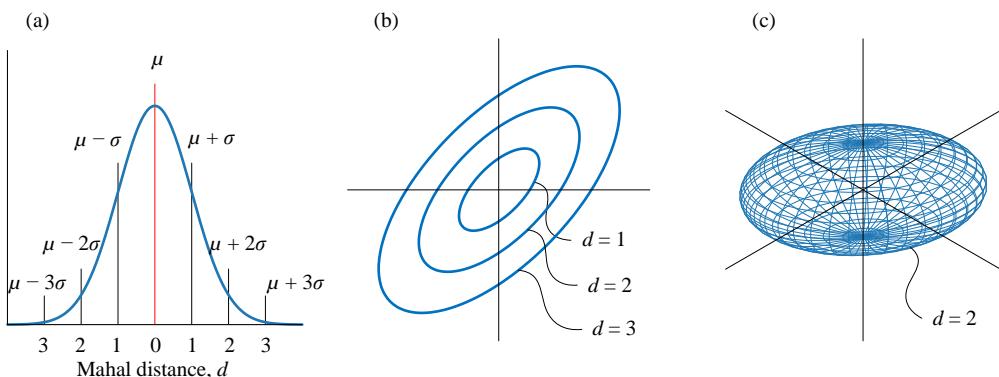


图 2. 马氏距离的几何形态

本章后文先比较三种常见距离：1) 欧氏距离；2) 标准化欧氏距离；3) 欧氏距离。

23.2 欧氏距离：最基本的距离

欧几里得距离 (Euclidean distance)，也称欧氏距离，是最“自然”的距离，是多维空间中两个点之间的绝对距离度量。

欧氏距离

x 和质心 μ 的欧氏距离定义为：

$$d = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu})} = \|\mathbf{x} - \boldsymbol{\mu}\| \quad (7)$$

欧氏距离本质上是 L^2 范数。

以鸢尾花花萼长度和花瓣长度两个特征数据为例，数据质心所在位置为：

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix} \quad (8)$$

注意，上式的两个特征单位为厘米。

如图 3 所示，平面上任意一点 x 到质心 μ 的欧氏距离的解析式为：

$$\begin{aligned} d &= \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{x} - \boldsymbol{\mu})} = \sqrt{\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix} \right)^T \left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix} \right)} \\ &= \sqrt{(x_1 - 5.843)^2 + (x_3 - 3.758)^2} \end{aligned} \quad (9)$$

图 3 所示的三个同心圆距离质心 μ 距离为 1 cm、2 cm、3 cm。此外，请大家注意图 4 中的网格，这个网格每个格子“方方正正”，边长都是 1 cm。

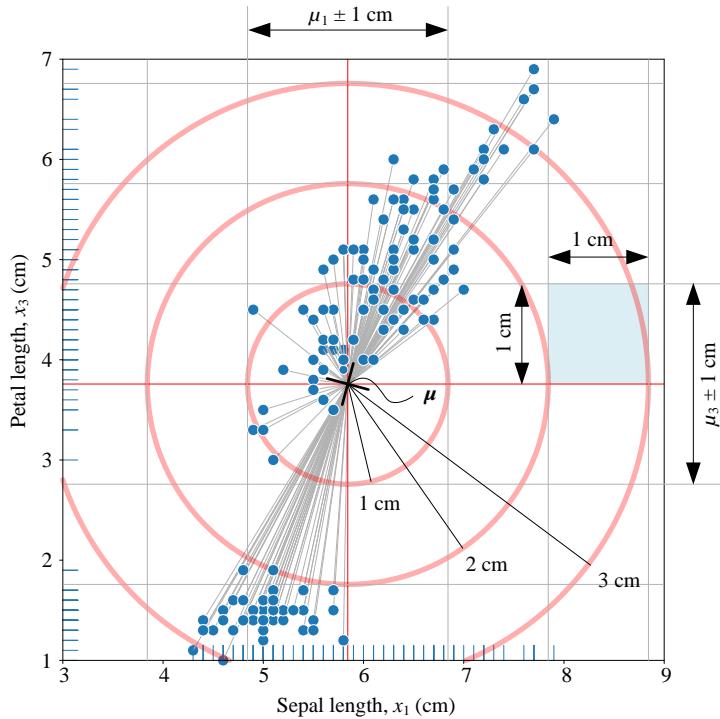


图 3. 花萼长度、花瓣长度平面上的欧氏距离等高线和网格

23.3 标准化欧氏距离：两个视角

第一视角：正椭圆

标准化欧氏距离 (standardized Euclidean distance) 定义如下：

$$d = \sqrt{(x - \mu)^T D^{-1} D^{-1} (x - \mu)} \quad (10)$$

其中， D 为对角方阵，对角线元素为标准差，运算如下：

$$D = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_D \end{bmatrix} \quad (11)$$

特别地，当 $D = 2$ 时，标准化欧氏距离为：

$$d = \sqrt{\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2}} = \sqrt{z_1^2 + z_2^2} \quad (12)$$

其中， z_1 和 z_3 是两个特征的 Z 分数。可以说， z_1 的单位是 σ_1 ， z_3 的单位是 σ_3 。

如图 3 所示， x_1x_3 平面上任意一点 x 到质心 μ 的标准化欧氏距离为：

$$d = \sqrt{\frac{(x_1 - 5.843)^2}{0.685} + \frac{(x_3 - 3.758)^2}{3.116}} \quad (13)$$

上式中，鸢尾花萼长度数据的方差为 0.685 cm^2 ，标准差 σ_1 为 0.827 cm 。花瓣长度数据的方差为 3.116 cm^2 ，标准差 σ_3 为 1.765 cm 。

图 4 所示为在花萼长度、花瓣长度平面上标准化欧氏距离为 1、2、3 的三个正椭圆。1、2、3 的单位可以理解为标准差。

大家注意图 4 中网格，网格的格子为矩形。矩形的宽度为 $\sigma_1 = 0.827 \text{ cm}$ ，矩形的长度为 $\sigma_3 = 1.765 \text{ cm}$ 。

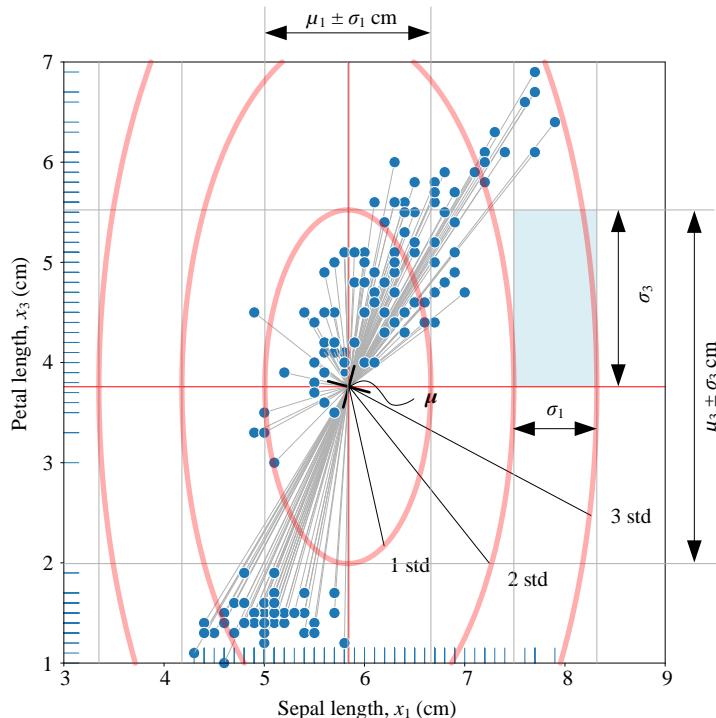


图 4. 花萼长度、花瓣长度平面上的标准化欧氏距离和网格

第二视角：正圆

先计算花萼长度、花瓣长度的 Z 分数 z_1 、 z_3 ：

$$z_1 = \frac{x_1 - 5.843}{0.827}, \quad z_3 = \frac{x_3 - 3.758}{1.765} \quad (14)$$

几何视角，上式经过了中心化、缩放两步。

然后再计算标准化欧氏距离：

$$d = \sqrt{z_1^2 + z_3^2} \quad (15)$$

图 5 所示花萼长度 z 分数、花瓣长度 z 分数平面上的标准化欧氏距离等高线。不难发现，在这个平面上，等高线为正圆，圆心位于原点。

图 5 中网格为正方形，这是因为数据已经标准化。

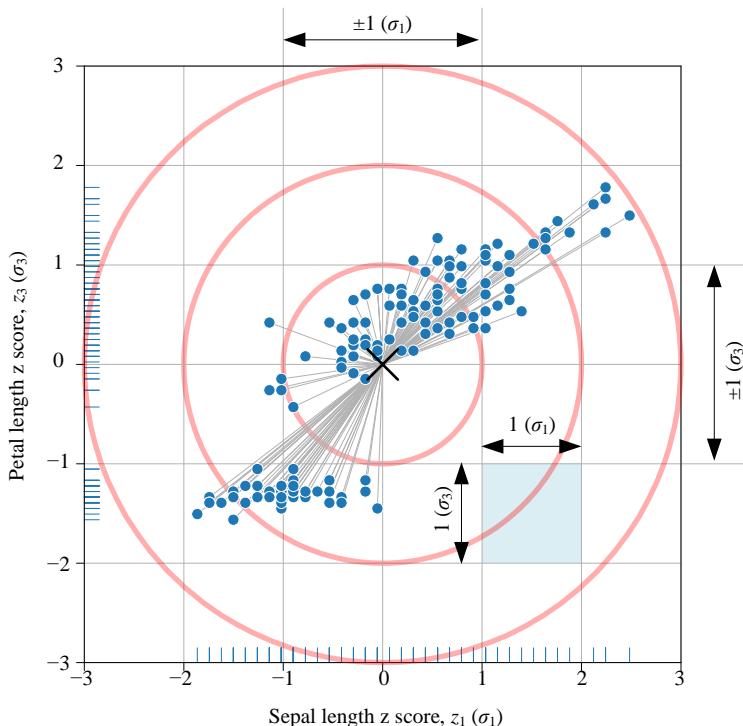


图 5. 花萼长度 z 分数、花瓣长度 z 分数平面上的标准化欧氏距离

23.4 马氏距离：两个视角

旋转椭圆

鸢尾花花萼长度、花瓣长度协方差矩阵 Σ 为：

$$\Sigma = \begin{bmatrix} 0.685 & 1.274 \\ 1.274 & 3.116 \end{bmatrix} \quad (16)$$

协方差 Σ 的逆为：

$$\Sigma^{-1} = \begin{bmatrix} 6.075 & -2.484 \\ -2.484 & 1.336 \end{bmatrix} \quad (17)$$

代入(1), 得到马氏距离的解析式：

$$\begin{aligned}
 d &= \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \begin{bmatrix} 6.075 & -2.484 \\ -2.484 & 1.336 \end{bmatrix} (\mathbf{x} - \boldsymbol{\mu})} \\
 &= \sqrt{\left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix} \right)^T \begin{bmatrix} 6.075 & -2.484 \\ -2.484 & 1.336 \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_3 \end{bmatrix} - \begin{bmatrix} 5.843 \\ 3.758 \end{bmatrix} \right)} \\
 &= \sqrt{6.08x_1^2 - 4.97x_1x_3 + 1.34x_3^2 - 52.32x_1 + 18.99x_3 + 117.21}
 \end{aligned} \tag{18}$$

图6中三个椭圆分别代表马氏距离为1、2、3。这个旋转椭圆的长轴就是第25章要介绍的**第一主成分** (first principal component) 方向, 而旋转椭圆的短轴就是**第二主成分** (second principal component) 方向。

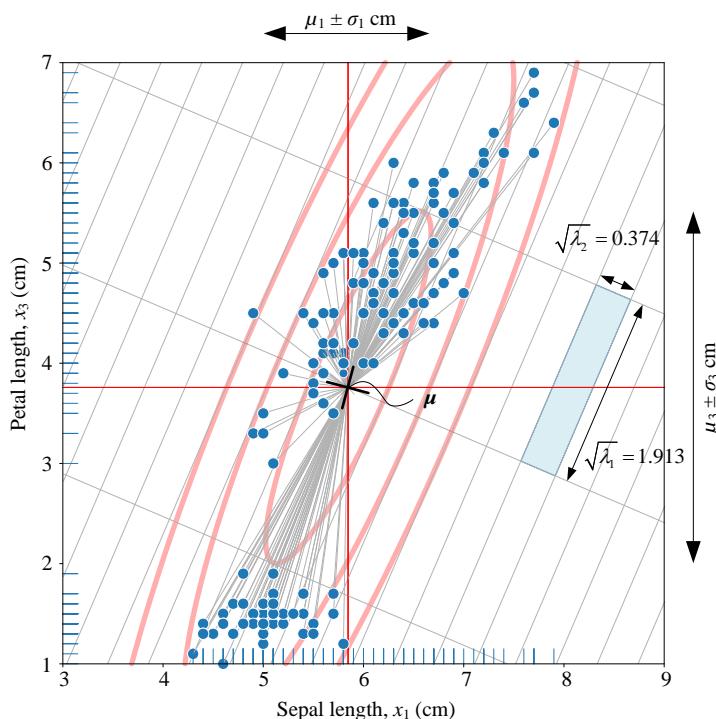


图6. 花萼长度、花瓣长度平面上的马氏距离等高线和网格

对协方差矩阵特征值分解得到的特征值方阵为：

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 3.661 \\ 0.140 \end{bmatrix} \tag{19}$$

两个特征值实际上就是数据投影在第一、第二主成分方向上的结果的方差, 也叫主成分方差。上式的单位也都是平方厘米 cm²。

而这两个特征值的平方根就是主成分标准差：

$$\sqrt{\lambda_1} = 1.913 \text{ cm}, \quad \sqrt{\lambda_2} = 0.374 \text{ cm} \quad (20)$$

它俩分别是旋转椭圆的半长轴、半短轴长度。

如图 6 所示，图中的网格就是度量马氏距离的坐标系。网格矩形倾斜角度和主成分方向相同。矩形的长度为 $\sqrt{\lambda_1}$ ，宽度为 $\sqrt{\lambda_2}$ 。

第二视角：正圆

令：

$$\mathbf{z} = \mathbf{A}^{\frac{-1}{2}} \mathbf{V}^T \begin{pmatrix} \mathbf{x} - \boldsymbol{\mu} \\ \text{Scale} \\ \text{Rotate} \\ \text{Centralize} \end{pmatrix} \quad (21)$$

将上式代入 (1)，得到马氏距离为 \mathbf{z} 的 L^2 范数：

$$d = \sqrt{\mathbf{z}^T \mathbf{z}} = \|\mathbf{z}\| \quad (22)$$

如图 7 所示，在第一、第二主成分平面上，马氏距离为正圆。

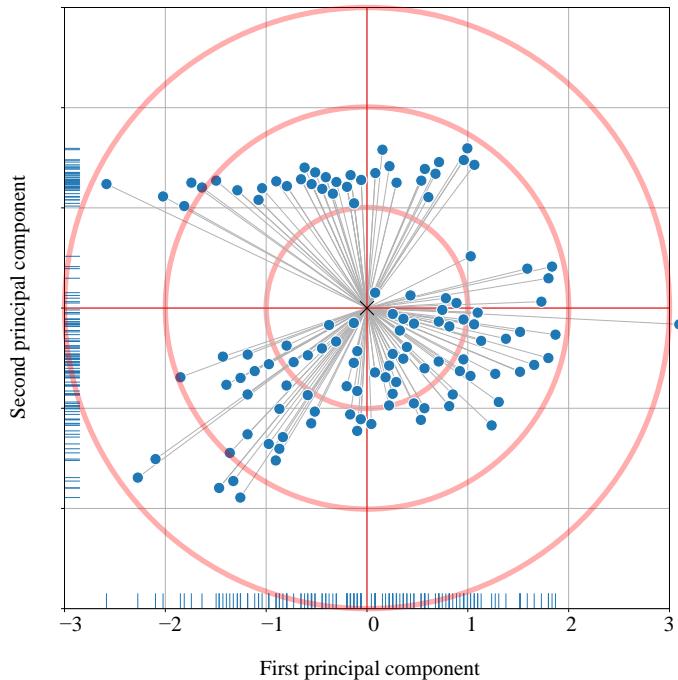
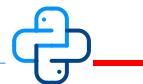


图 7. 第一、第二主成分平面上马氏距离等高线和网格



Bk5_Ch23_01.py 绘制图 3、图 4、图 6。

成对特征图

马氏距离椭圆也可以画在成对特征图上。图 8 和图 9 分别展示考虑不考虑标签的马氏距离椭圆。这些图像可以帮助我们分析理解数据，比如解读相关性、发现离群值等。



《数据有道》一册将专门讲解如何发现离群值。

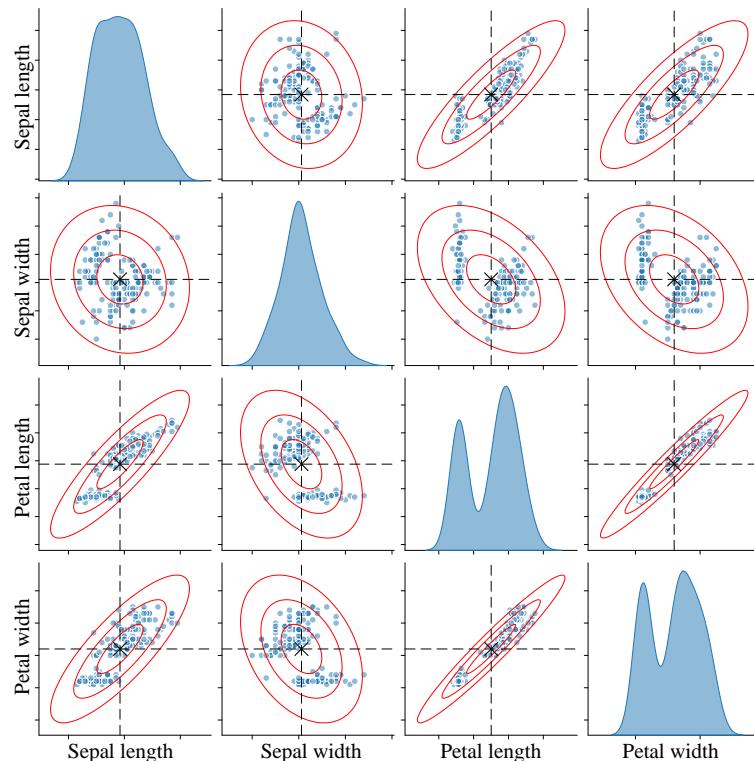


图 8. 成对特征图上绘制马氏距离等高线，不考虑标签

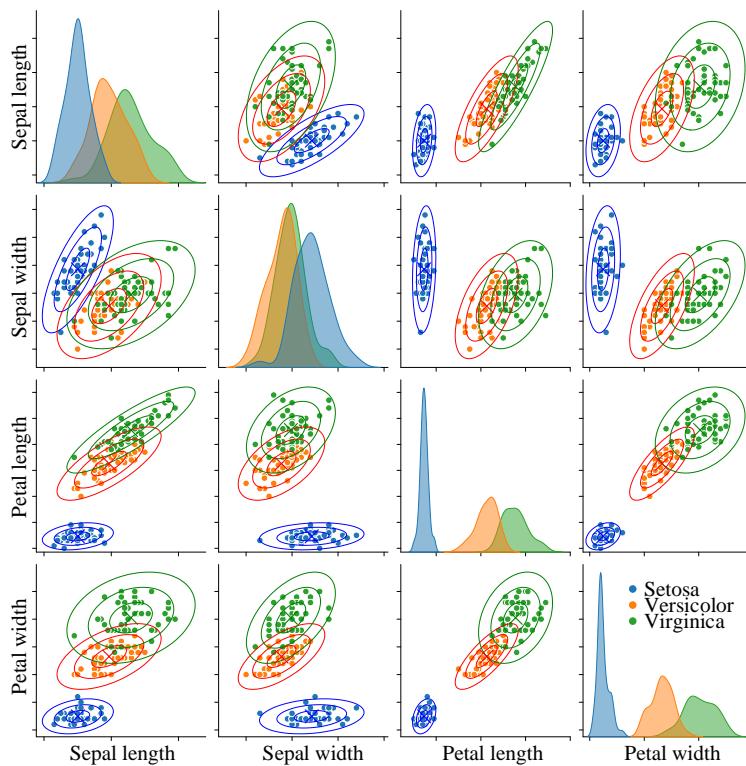
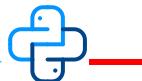


图 9. 成对特征图上绘制马氏距离等高线，考虑标签



Bk5_Ch23_02.py 绘制图 8 和图 9。

23.5 马氏距离和卡方分布

本书第 9 章介绍过一元高斯分布的“68-95-99.7 法则”。这个法则具体是指，如果数据近似服从一元高斯分布 $N(\mu, \sigma)$ ，则约 68.3%、95.4% 和 99.7% 的数据分布在距均值 (μ) 1 个 $(\mu \pm \sigma)$ 、2 个 $(\mu \pm 2\sigma)$ 和 3 个 $(\mu \pm 3\sigma)$ 正负标准差范围之内。

而 68.3%、95.4% 和 99.7% 这三个数实际上卡方分布直接相关。当 $D = 1$ 时， X_i 服从正态分布 $N(\mu_1, \sigma_1)$ ，经过标准化得到的随机变量 Z_i 则服从标准正态分布：

$$Z_i = \frac{X_i - \mu_1}{\sigma_1} \sim N(0, 1) \quad (23)$$

也就是说， Z_i 的平方服从自由度为 1 的卡方分布：

$$Z_i^2 \sim \chi_{(df=1)}^2 \quad (24)$$

⚠ 注意，实际上 Z_1 的平方再开方，即 Z_1 的绝对值，就是马氏距离。

$D = 2$ 时，马氏距离平方 d^2 服从 $df = 2$ 的卡方分布：

$$d^2 \sim \chi_{(df=2)}^2 \quad (25)$$

D 维马氏距离的平方则服从自由度为 D 的卡方分布：

$$d^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \sim \chi_{(df=D)}^2 \quad (26)$$

也就是说，距离为 d 的马氏距离超椭圆围成的几何图形内部的概率 α 可以用卡方分布 CDF 查表获得。

比如，SciPy 中卡方分布的对象为 `scipy.stats.distributions.chi2`，计算 $D = 2$ ，马氏距离 $d = 3$ 条件下，马氏距离椭圆围成的图形的概率 α 为 `scipy.stats.distributions.chi2.cdf(d^2 = 9, df = 2)`。

这实际上也回答了本书第 10 章的问题，具体如图 10 所示。请大家查表回答这个问题。

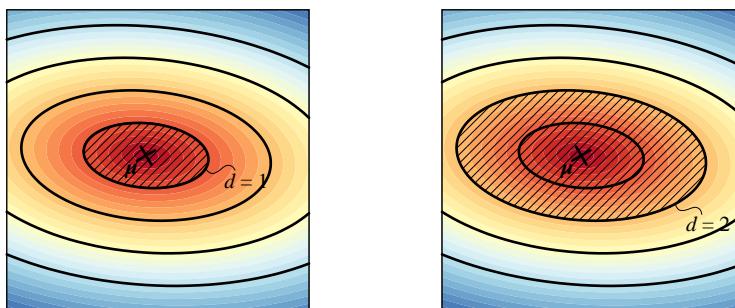


图 10. 求阴影区域对应的概率，来自本书第 10 章

相反，如果给定概率值 α 和自由度，可以用卡方分布的百分点函数 PPF，即 CDF 的逆函数 (inverse CDF)，反求马氏距离的平方 d^2 。这个值开方就是马氏距离 d 。

比如，给定概率值 0.9，自由度为 2，利用 `scipy.stats.distributions.chi2.ppf(0.9, df=2)` 可以求得马氏距离的平方值 d^2 ，开方就是马氏距离 d 。

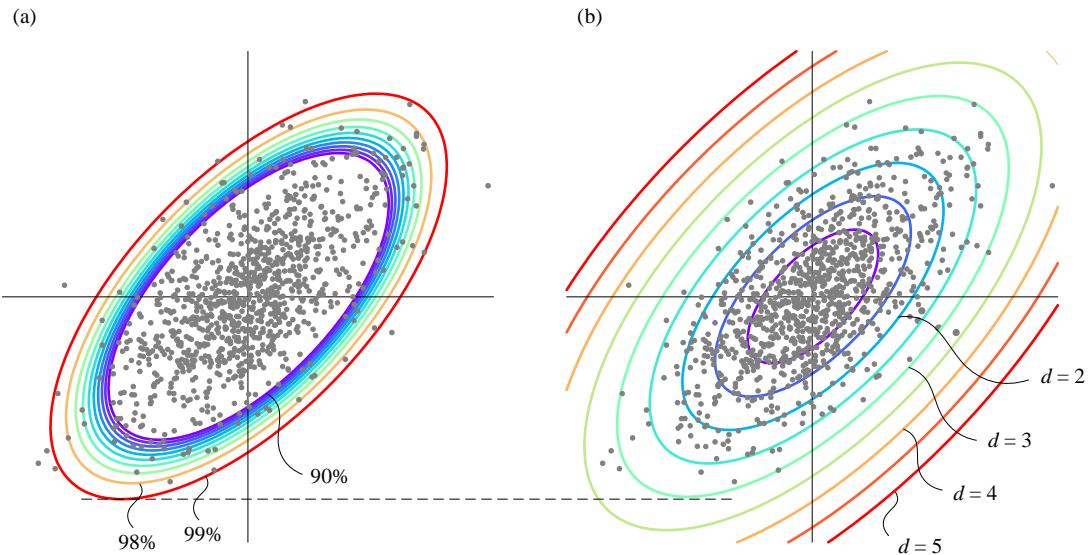
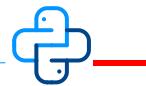
如图 11 (a) 所示，自由度为 2，给定一系列概率值 (0.90 ~ 0.99)，利用卡方分布的百分点函数 PPF，我们便获得一系列马氏距离椭圆。图 11 (b) 对照马氏距离取值为 1 ~ 5。

这些椭圆中，马氏距离 3 几乎对应 99% 这个概率值。也就是说，如果二元随机数近似服从二元高斯分布，约有 99% 的随机数落在马氏距离为 3 的椭圆内。



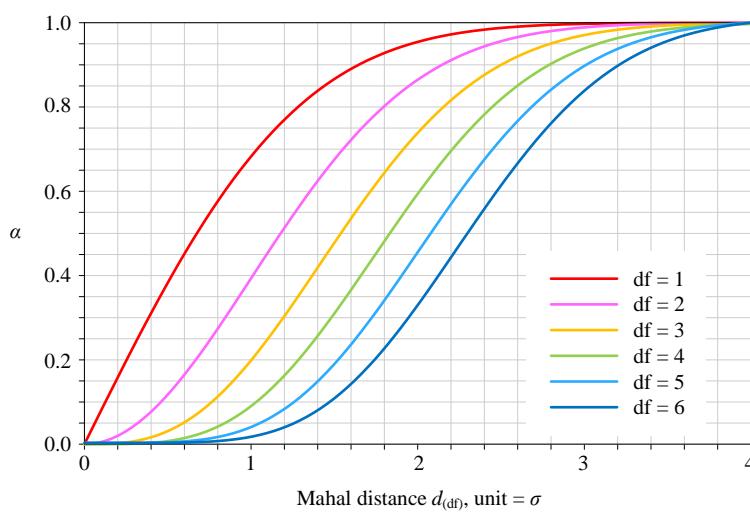
用卡方分布将马氏距离转换为概率时，有些文献错误地将自由度给定为 $D - 1$ ，即特征数 D 减 1。下面这篇文章详尽地解释如何正确设定自由度，建议大家参考。

<https://peerj.com/articles/6678/>

图 11 特征数 $D = 2$ 时，概率值 α 和马氏距离椭圆位置

Bk5_Ch23_03.py 绘制图 11。

图 12 所示为马氏距离 d 、自由度 df 、概率值 α 三者关系曲线。

图 12 马氏距离 d 、自由度 df 、概率值 α 三者关系

为了方便查表，大家可以参考图 13 和图 14。图 13 中，给定马氏距离 d 、自由度 df ，查表得到 α 。这张表中，我们可以看到一元高斯分布的 68-95-99.7 法则。

而自由度 $df = 2$ 时，这个法则变为马氏距离为 1、2、3 的椭圆对应 39%、86%、98.9%，我们也可以管它叫 39-86-98.9 法则。

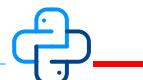
图 14 中，给定概率值 α 、自由度 df ，查表得到马氏距离 d 。

Degree of freedom, df	Mahal distance, d												
	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3	3.25	3.5	3.75	4
1	0.6827	0.7887	0.8664	0.9199	0.9545	0.9756	0.9876	0.9940	0.9973	0.9988	0.9995	0.9998	0.9999
2	0.3935	0.5422	0.6753	0.7837	0.8647	0.9204	0.9561	0.9772	0.9889	0.9949	0.9978	0.9991	0.9997
3	0.1987	0.3321	0.4778	0.6179	0.7385	0.8327	0.8999	0.9440	0.9707	0.9857	0.9934	0.9972	0.9989
4	0.0902	0.1845	0.3101	0.4526	0.5940	0.7191	0.8188	0.8910	0.9389	0.9681	0.9844	0.9929	0.9970
5	0.0374	0.0943	0.1864	0.3096	0.4506	0.5917	0.7174	0.8179	0.8909	0.9392	0.9685	0.9848	0.9932
6	0.0144	0.0448	0.1047	0.1990	0.3233	0.4642	0.6042	0.7281	0.8264	0.8971	0.9434	0.9711	0.9862

图 13. 给定马氏距离 d 、自由度 df ，查表得到概率值 α

Degree of freedom, df	Probability α that the random value will fall inside the ellipsoid												
	0.9	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	0.993	0.996	0.999
1	1.6449	1.6954	1.7507	1.8119	1.8808	1.9600	2.0537	2.1701	2.3263	2.5758	2.6968	2.8782	3.2905
2	2.1460	2.1945	2.2475	2.3062	2.3721	2.4477	2.5373	2.6482	2.7971	3.0349	3.1502	3.3231	3.7169
3	2.5003	2.5478	2.5997	2.6571	2.7216	2.7955	2.8829	2.9912	3.1365	3.3682	3.4806	3.6492	4.0331
4	2.7892	2.8361	2.8873	2.9439	3.0074	3.0802	3.1663	3.2729	3.4158	3.6437	3.7542	3.9199	4.2973
5	3.0391	3.0856	3.1363	3.1923	3.2552	3.3272	3.4124	3.5178	3.6590	3.8841	3.9932	4.1568	4.5293
6	3.2626	3.3088	3.3591	3.4147	3.4770	3.5485	3.6329	3.7373	3.8773	4.1002	4.2083	4.3702	4.7390

图 14. 给定概率值 α 、自由度 df ，查表得到马氏距离 d



Bk5_Ch23_04.py 绘制图 12。



马氏距离是一种基于统计学的距离度量方法，用于衡量两个样本之间的相似度或距离。马氏距离考虑了各个特征之间的相关性，相比于欧氏距离或曼哈顿距离等传统距离度量方法，更适合用于高维数据集合。马氏距离被广泛应用于分类、聚类、异常检测等领域，特别是在高维数据集合的分析和处理中。由于它考虑了各个特征之间的相关性，因此在某些情况下比传统距离度量方法更为有效和准确。

《统计至简》一册中椭圆无处不在，希望大家日后看到椭圆，就能想到协方差矩阵、多元高斯分布、相关性、旋转、缩放、特征值分解、置信区间、离群值、马氏距离、线性回归、主成分分析等等内容。更能“看到”日月所属、天体运转、星辰大海。



Linear Regression

线性回归

以概率统计、几何、矩阵分解、优化为视角



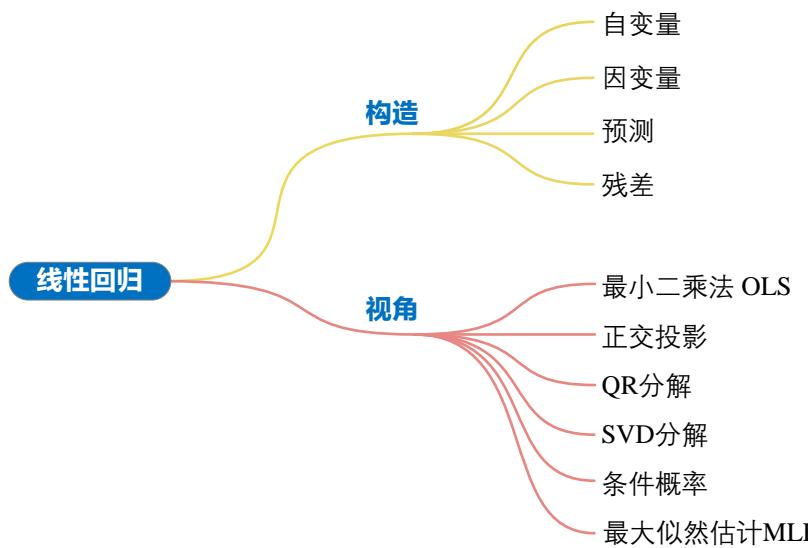
我们必须承认，有多少数字，就有多少正方形。

We must say that there are as many squares as there are numbers.

——伽利略·伽利莱 (Galilei Galileo) | 意大利物理学家、数学家及哲学家 | 1564 ~ 1642



- ◀ `matplotlib.pyplot.quiver()` 绘制箭头图
- ◀ `numpy.cov()` 计算协方差矩阵
- ◀ `seaborn.heatmap()` 绘制热图
- ◀ `seaborn.jointplot()` 绘制联合分布/散点图和边际分布
- ◀ `seaborn.kdeplot()` 绘制 KDE 核概率密度估计曲线
- ◀ `seaborn.pairplot()` 绘制成对分析图
- ◀ `statsmodels.api.add_constant()` 线性回归增加一列常数 1
- ◀ `statsmodels.api.OLS()` 最小二乘法函数



24.1 再聊线性回归

线性回归 (linear regression) 是最为常用的回归建模技术。它是利用线性关系建立因变量与一个或多个自变量之间的联系。线性回归模型相对简单，可解释性强，应用广泛。

鸢尾花书从不同视角介绍过线性回归。比如，《数学要素》从代数、几何、优化角度讲过线性回归，《矩阵力量》则从线性代数、正交投影、矩阵分解视角分析线性回归。本章一方面总结这几个视角，另外一方面以条件概率、MLE 为视角再谈线性回归。

有监督学习

《矩阵力量》一书提过，线性回归是一种**有监督学习** (supervised learning)。有监督学习是一种机器学习方法，它利用已知的标签或输出值来训练模型，并用于预测未知的标签或输出值。在有监督学习中，我们通常会提供一组已知的训练样本，每个样本都包含一组输入特征和相应的输出标签。模型通过分析这些训练样本来学习如何将输入特征映射到输出标签，从而能够用于预测未知的输出值。

有监督学习通常分为两个主要的子类别：**分类** (classification) 和**回归** (regression)。在分类问题中，目标是将输入特征映射到有限的离散类别。在回归问题中，目标是将输入特征映射到连续的输出值。



《数据有道》将介绍更多有关回归算法，而《机器学习》将关注常见分类算法。

简单线性回归

简单线性回归 (Simple Linear Regression, SLR) 也叫**一元线性回归** (univariate linear regression)，是指模型中只含有一个自变量和一个因变量，表达式如下：

$$y = \underbrace{b_0 + b_1 x}_{\hat{y}} + \varepsilon \quad (1)$$

其中， b_0 为**截距项** (intercept)， b_1 代表**斜率** (slope)。

x 又常被称作**自变量** (independent variable)、**解释变量** (explanatory variable) 或**回归元** (regressor)、**外生变量** (exogenous variables)、**预测变量** (predictor variables)；

y 常被称作**因变量** (dependent variable)、**被解释变量** (explained variable)、或**回归子** (regressand)、**内生变量** (endogenous variable)、**响应变量** (response variable) 等。

ε 为**残差项** (residuals)、**误差项** (error term)、**干扰项** (disturbance term) 或**噪音项** (noise term)。

图 1 所示为平面上的一个线性回归关系。

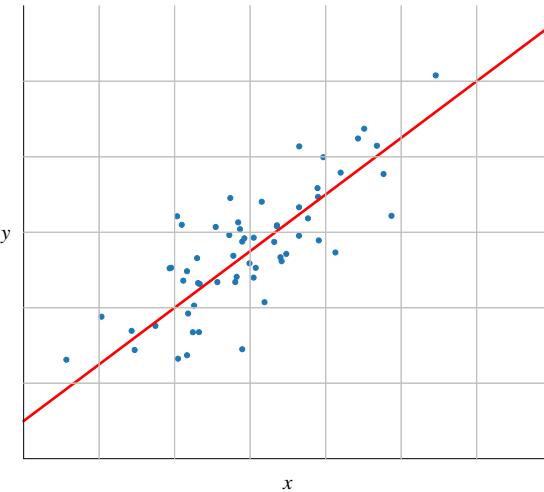


图 1. 平面上，一元线性回归

预测

利用 (1) 做预测，预测值 \hat{y} 为：

$$\hat{y} = b_0 + b_1 x \quad (2)$$

⚠ 注意，“戴帽子”的 \hat{y} 表示预测值。

(2) 对应图 1 中的红色直线。

对于第 i 个数据点，预测值 $\hat{y}^{(i)}$ 可以通过下式计算得到：

$$\hat{y}^{(i)} = b_0 + b_1 x^{(i)} \quad (3)$$

残差

(1) 中残差项为：

$$\varepsilon = y - (b_0 + b_1 x) = y - \hat{y} \quad (4)$$

如图 2 所示，在平面上，残差项是 y 和 \hat{y} 之间的纵轴上的高度差。

⚠ 注意，平面上，线性回归和主成分分析的结果看上去都是一条直线。但是两者差距甚远。线性回归是有监督学习，而主成分分析是无监督学习。从距离角度来看，线性回归关注的是沿纵轴的高度差，而主成分分析则是聚焦点到直线的距离。此外，从椭圆的角度来看，主成分分析对应椭圆的长轴、短轴，而线性回归则和与椭圆相切的矩形有关。“鸢尾花书”《编程不难》聊过这个话题，请大家回顾。

真实观察值 $y^{(i)}$ 和预测值 $\hat{y}^{(i)}$ 之差为第 i 个数据点的残差：

$$\varepsilon^{(i)} = y^{(i)} - \hat{y}^{(i)} \quad (5)$$

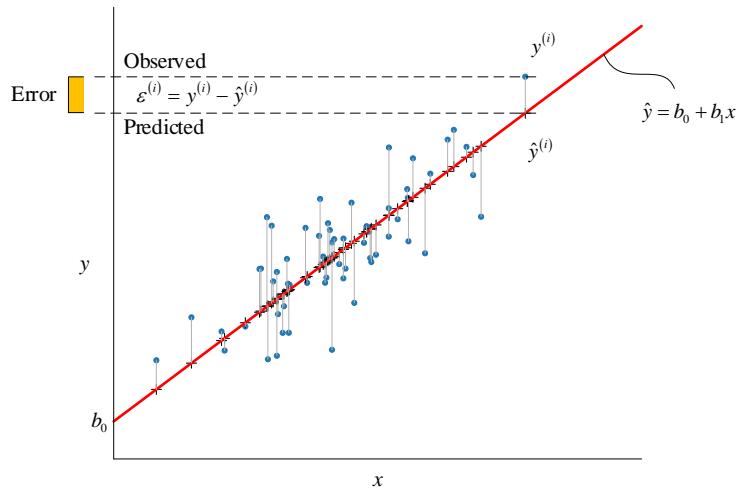


图 2. 简单线性回归中的残差项

矩阵形式

使用阵运算表达一元线性回归：

$$\mathbf{y} = b_0 \mathbf{I} + b_1 \mathbf{x} + \boldsymbol{\varepsilon} \quad (6)$$

\mathbf{I} 为和 \mathbf{x} 形状相同的全 1 列向量；自变量数据 \mathbf{x} 、因变量数据 \mathbf{y} 和残差项 $\boldsymbol{\varepsilon}$ 分别包括 n 个样本对应的列向量为：

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix} \quad (7)$$

图 3 解释 (6) 给出的矩阵运算。

$$\begin{array}{ccccccccc} \mathbf{y} & = & b_0 \mathbf{I} & + & b_1 \mathbf{x} & + & \boldsymbol{\varepsilon} \\ \begin{array}{c} \text{red} \\ \text{red} \\ \text{red} \\ \text{red} \end{array} & & \begin{array}{c} \text{grey} \\ \text{grey} \\ \text{grey} \\ \text{grey} \end{array} & & \begin{array}{c} \text{blue} \\ \text{blue} \\ \text{blue} \\ \text{blue} \end{array} & & \begin{array}{c} \text{yellow} \\ \text{yellow} \\ \text{yellow} \\ \text{yellow} \end{array} \\ n \times 1 & & n \times 1 & & n \times 1 & & n \times 1 \end{array}$$

图 3. 用矩阵运算表达一元回归

预测值构成的列向量 \hat{y} 为：

$$\hat{y} = b_0 \mathbf{I} + b_1 \mathbf{x} \quad (8)$$

如图 4 所示， \hat{y} 是 \mathbf{I} 和 \mathbf{x} 的线性组合。

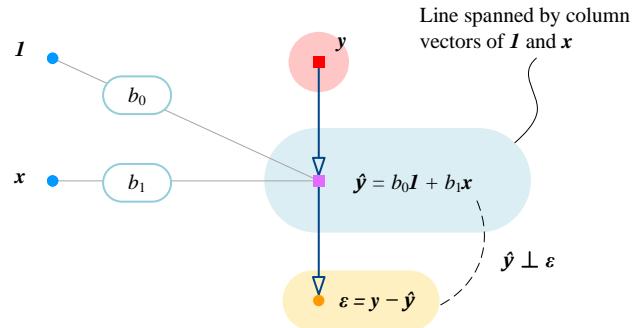


图 4. 一元最小二乘法线性回归数据关系

残差项列向量 ϵ 为：

$$\epsilon = y - \hat{y} \quad (9)$$

图 5 可可视化求解残差项列向量 ϵ 过程。

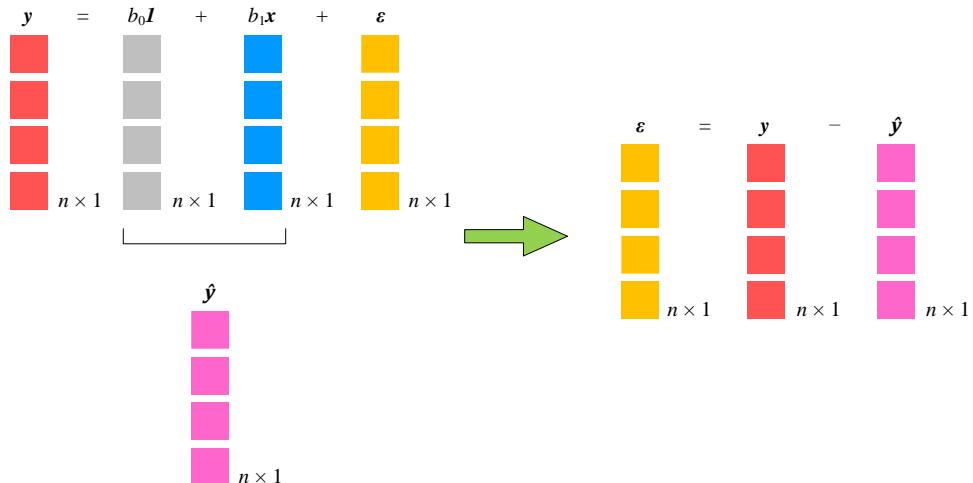


图 5. 求解残差项列向量

问题来了，如何确定参数 b_0 、 b_1 ？

24.2 最小二乘法

最小二乘法 (ordinary least squares, OLS) 通过**最小化残差值平方和** (sum of squared estimate of errors, SSE), 来计算得到最佳的拟合回归线参数:

$$\arg \min_{b_0, b_1} \text{SSE} = \arg \min_{b_0, b_1} \sum_{i=1}^n (\varepsilon^{(i)})^2 \quad (10)$$

残差平方和 SSE 为:

$$\text{SSE} = \sum_{i=1}^n (\varepsilon^{(i)})^2 = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 \quad (11)$$

⚠ 注意, “鸢尾花书”用 SSE 表达残差值平方和; 也有很多文献使用 RSS (residual sum of squares) 代表残差值平方和。

从几何角度, 图 6 中的每一个正方形的边长为 $\varepsilon^{(i)}$, 该正方形的面积代表一个残差平方项 $(\varepsilon^{(i)})^2$; 图 6 所有正方形面积之和便是残差平方和 SSE。

→ 我们在《数学要素》第 24 章聊过残差平方和 SSE 可以写成一个二元函数 $f(b_0, b_1)$ 。 $f(b_0, b_1)$ 对应的图像如图 7 所示。

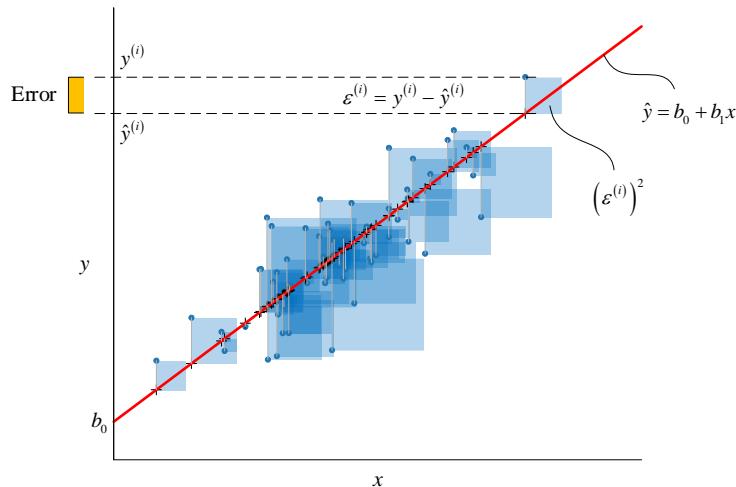
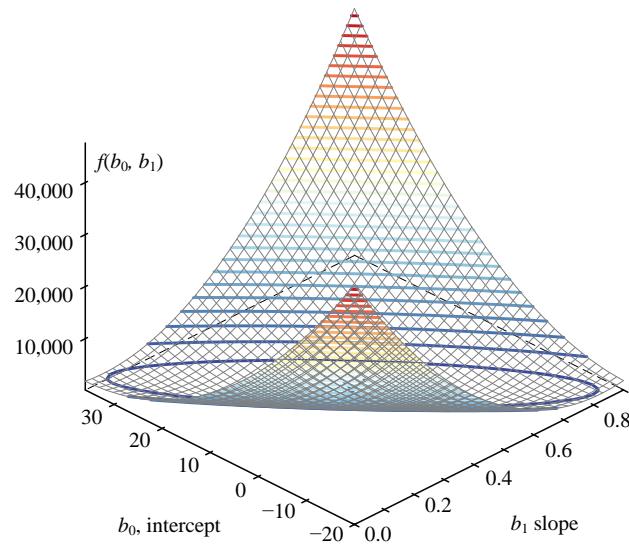


图 6. 残差平方和的几何意义

图 7. 误差平方和 SSE 随 b_0 、 b_1 变化构造的开口向上抛物曲面，图片来自《数学要素》第 24 章

24.3 优化问题

用线性代数工具构造 OLS 优化问题：

$$\arg \min_b \|y - Xb\| \quad (12)$$

也可以写成：

$$\arg \min_b \|\varepsilon\|^2 = \varepsilon^T \varepsilon \quad (13)$$

令

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \quad X = [I \quad x] \quad (14)$$

其中， X 又叫**设计矩阵** (design matrix)。

\hat{y} 可以写成：

$$\hat{y} = Xb \quad (15)$$

残差向量 ε 可以写成：

$$\varepsilon = y - b_0 I - b_1 x = y - Xb \quad (16)$$

定义 $f(b)$ 为：

$$f(b) = \varepsilon^T \varepsilon = (y - Xb)^T (y - Xb) \quad (17)$$

$f(\mathbf{b})$ 对 \mathbf{b} 求一阶导为 $\mathbf{0}$ 得到等式：

$$\frac{\partial f(\mathbf{b})}{\partial \mathbf{b}} = 2\mathbf{X}^T \mathbf{X}\mathbf{b} - 2\mathbf{X}^T \mathbf{y} = \mathbf{0} \quad (18)$$

如果 $\mathbf{X}^T \mathbf{X}$ 可逆，则 \mathbf{b} 为：

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (19)$$

24.4 投影视角

《矩阵力量》一本特别强调 OLS 的投影视角。

如图 8 所示，在 \mathbf{I} 和 \mathbf{x} 撑起平面 H 上，向量 \mathbf{y} 的投影为 $\hat{\mathbf{y}}$ ，而残差 ε 垂直于这个平面：

$$\begin{aligned} \varepsilon \perp \mathbf{I} &\Rightarrow \mathbf{I}^T \varepsilon = 0 \Rightarrow \mathbf{I}^T (\mathbf{y} - \hat{\mathbf{y}}) = 0 \\ \varepsilon \perp \mathbf{x} &\Rightarrow \mathbf{x}^T \varepsilon = 0 \Rightarrow \mathbf{x}^T (\mathbf{y} - \hat{\mathbf{y}}) = 0 \end{aligned} \quad (20)$$

以上两式合并：

$$\underbrace{[\mathbf{I} \quad \mathbf{x}]}_{\mathbf{X}}^T (\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0} \quad (21)$$

整理得到：

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X}\mathbf{b} \quad (22)$$

这和 (18) 一致。

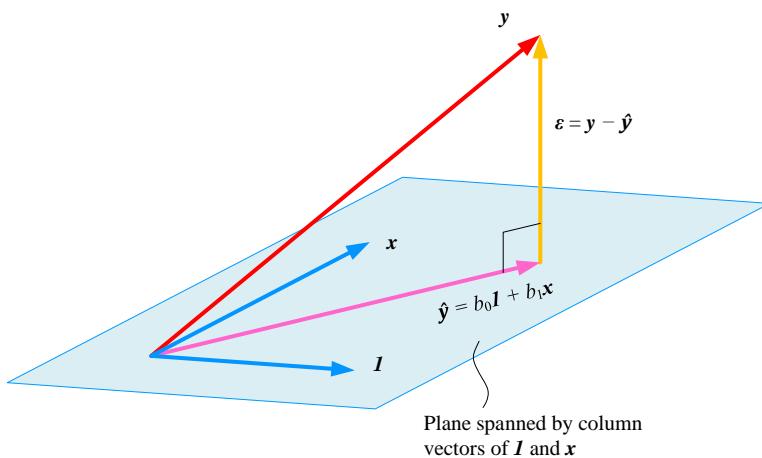


图 8. 几何角度解释一元最小二乘结果，二维平面

24.5 线性方程组：代数视角

实际上，下式就是一个超定方程组 (overdetermined system)：

$$\mathbf{y} = \mathbf{X}\mathbf{b} \quad (23)$$

QR 分解

对 \mathbf{X} 进行 QR 分解得到：

$$\mathbf{X} = \mathbf{Q}\mathbf{R} \quad (24)$$

这样求得 \mathbf{b} 为：

$$\mathbf{b} = \mathbf{R}^{-1}\mathbf{Q}^T\mathbf{y} \quad (25)$$

奇异值分解

对 \mathbf{X} 进行完全型 SVD 分解得到：

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T \quad (26)$$

这样求得 \mathbf{b} 为：

$$\mathbf{b} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T\mathbf{y} \quad (27)$$

 《矩阵力量》一册介绍过 $\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T$ 是 \mathbf{X} 的摩尔-彭若斯广义逆 (Moore-Penrose inverse)。
 \mathbf{S}^{-1} 的主对角线非零元素为 \mathbf{S} 的非零奇异值倒数， \mathbf{S}^{-1} 其余对角线元素均为 0。

24.6 条件概率

条件期望



本书第 12 章介绍过，线性回归还可以从条件概率视角来看。

如果随机变量 (X, Y) 服从二元高斯分布，给定 $X = x$ 条件下， Y 的条件期望为：

$$\mu_{Y|X=x} = \text{cov}(X, Y) (\sigma_x^2)^{-1} (x - \mu_x) + \mu_Y = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} (x - \mu_x) + \mu_Y \quad (28)$$

这条回归直线的斜率为 $\rho_{X,Y}\sigma_Y/\sigma_X$, 且通过点 (μ_x, μ_y) 。

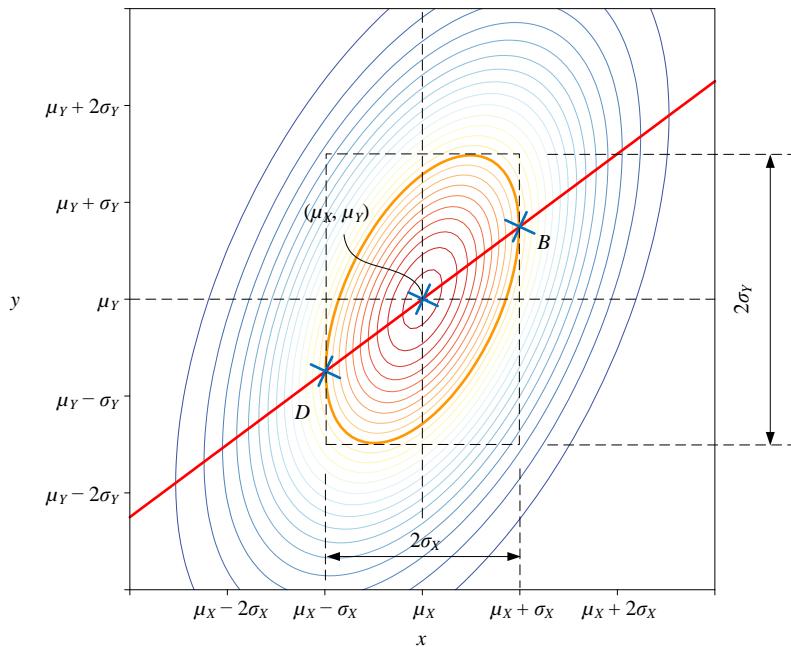
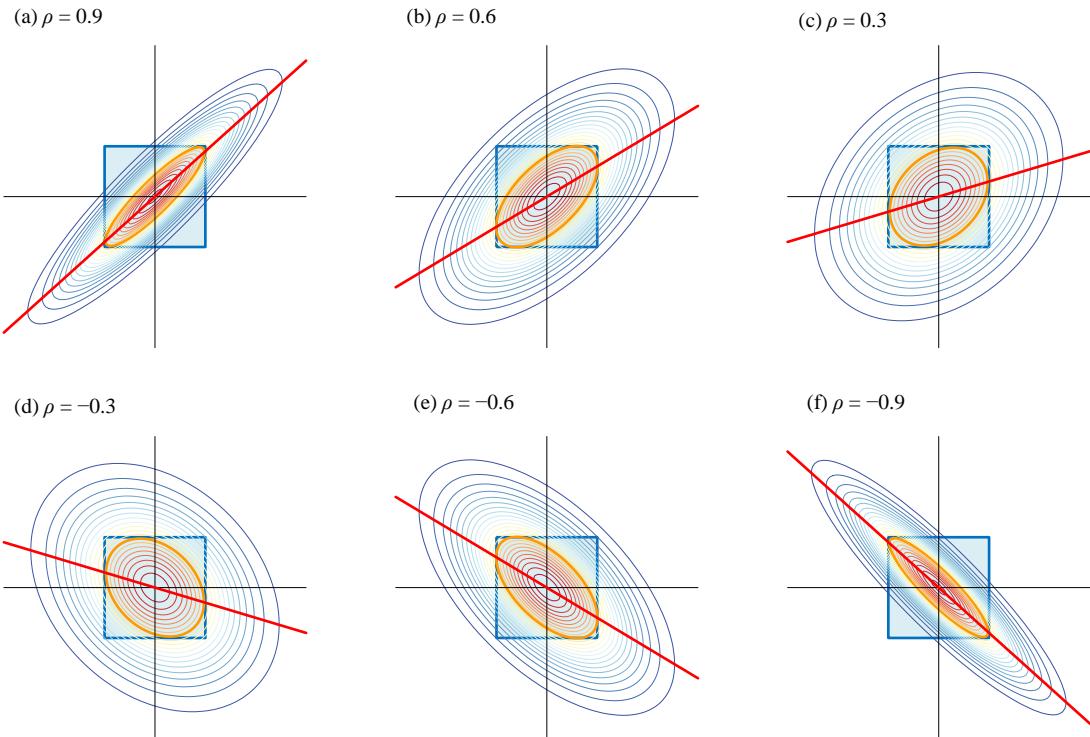


图 9. 给定 $X=x$ 的条件期望

图 10 所示为不同相关性系数条件下，回归直线和椭圆关系。

图 10. 条件期望直线位置和相关性系数关系, $\sigma_x = \sigma_y$

以鸢尾花为例

定义鸢尾花花瓣长度为 x , 鸢尾花花萼宽度为 y 。鸢尾花样本数据, x 和 y 的关系为:

$$y = 3.758 + 1.858 \begin{pmatrix} x - 5.843 \\ \rho_{x,y} \frac{\sigma_y}{\sigma_x} \end{pmatrix} \quad (29)$$

图 11 中散点为样本数据, 其中直线代表花瓣长度、花萼长度之间回归关系。这幅图中, 我们还绘制了马氏距离为 1 的椭圆。这个椭圆代表了花瓣长度、花萼长度的协方差矩阵。

图 12 所示为不考虑标签情况下, 鸢尾花的成对特征图以及特征之间的回归关系。图 13 所示为考虑标签情况下, 鸢尾花的成对特征图以及特征之间的回归关系。

⚠ 特别值得注意的是, 两个随机变量之间的线性回归关系不代表两者存在“因果关系”。

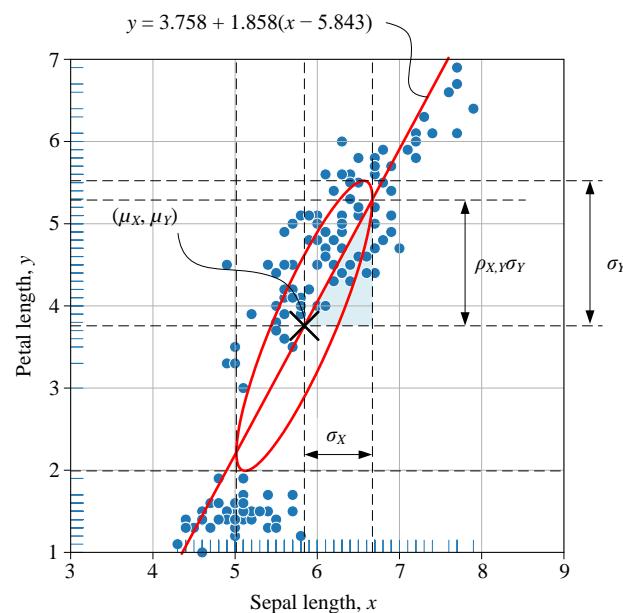
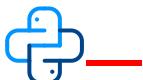
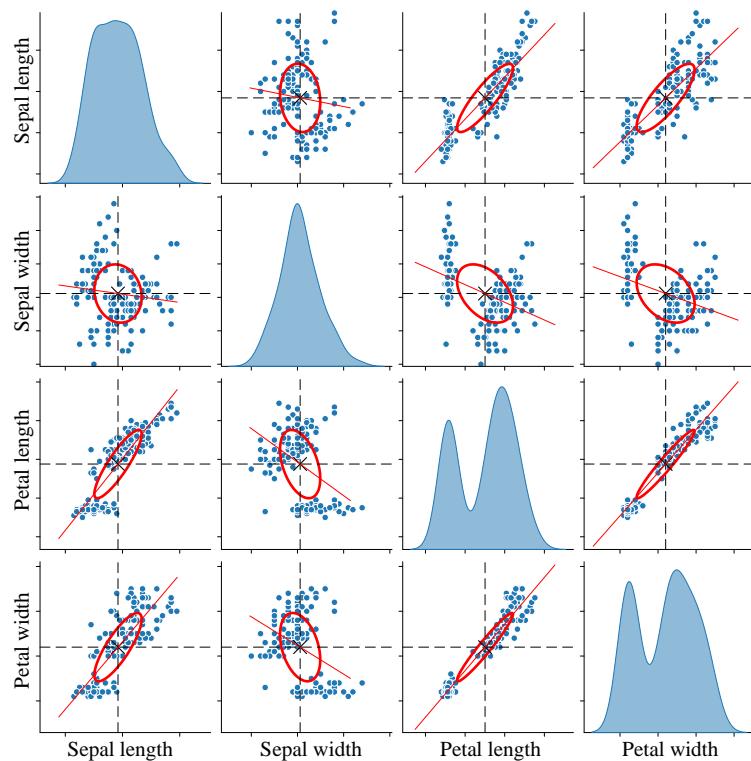


图 11. 花瓣长度、花萼长度之间回归关系



Bk5_Ch24_01.py 绘制图 11。



本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。
版权归清华大学出版社所有，请勿商用，引用请注明出处。

代码及 PDF 文件下载：<https://github.com/Visualize-ML>

本书配套微课视频均发布在 B 站——生姜 DrGinger：<https://space.bilibili.com/513194466>

欢迎大家批评指教，本书专属邮箱：jiang.visualize.ml@gmail.com

图 12. 成对特征图和回归关系

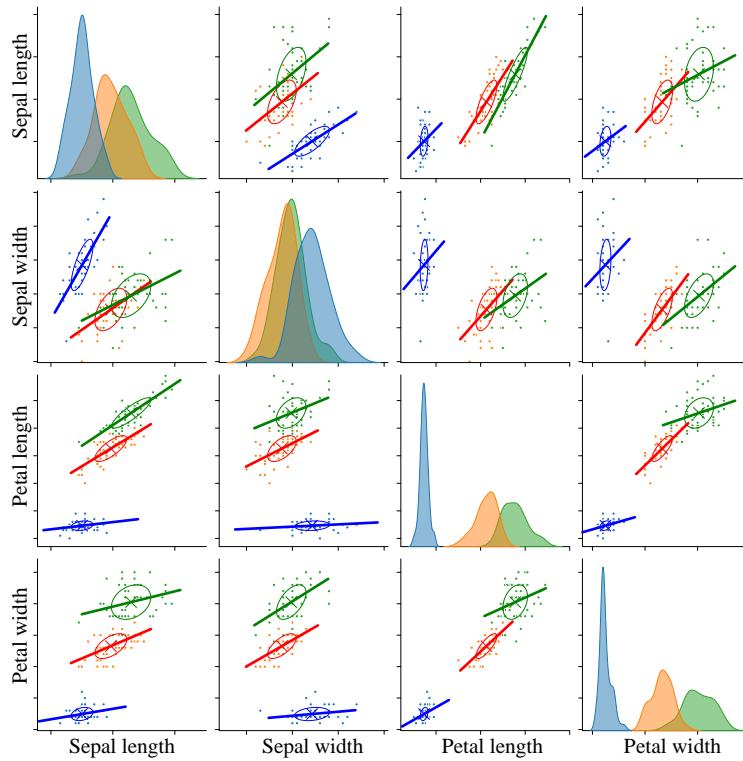


图 13. 成对特征图和回归关系，考虑分类标签



Bk5_Ch24_02.py 绘制图 12 和图 13。

24.7 最大似然估计 MLE

为了方便和本书前文有关最大似然估计内容对照阅读，本节中，线性回归解析式改写成：

$$y = \underbrace{\theta_0 + \theta_1 x}_{\hat{y}} + \varepsilon \quad (30)$$

对应的超定方程组写成：

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} \quad (31)$$

残差向量 ε 为：

$$\varepsilon = \mathbf{y} - \mathbf{X}\boldsymbol{\theta} \quad (32)$$

假设残差项服从正态分布：

$$\varepsilon \sim N(0, \sigma^2) \quad (33)$$

根据 (4)，也就是说 Y_i 服从：

$$Y_i \sim N(\theta_1 X_i + \theta_0, \sigma^2) \quad (34)$$

Y_i 的概率密度函数为：

$$f_{Y_i}(y_i; \theta_1 x_i + \theta_0, \sigma) = \frac{\exp\left(-\frac{(y_i - (\theta_1 x_i + \theta_0))^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \quad (35)$$

似然函数可以写成：

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \frac{\exp\left(-\frac{(y_i - (\theta_1 x_i + \theta_0))^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma} \quad (36)$$

对数似然函数为：

$$\ln L(\theta_0, \theta_1) = -n \ln(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^n (y_i - (\theta_1 x_i + \theta_0))^2}{2\sigma^2} \quad (37)$$

假设 σ 已知，最大化对数似然函数，等价于最小化 $\sum_{i=1}^n (y_i - (\theta_1 x_i + \theta_0))^2$ ，这和 (13) 优化问题一致。

$$\begin{aligned} \hat{\theta}_1 &= \frac{\sum_{i=1}^n (x^{(i)} - \mu_x)(y^{(i)} - \mu_y)}{\sum_{i=1}^n (x^{(i)} - \mu_x)^2} \\ \hat{\theta}_0 &= \mu_y - \hat{\theta}_1 \mu_x \end{aligned} \quad (38)$$

矩阵运算

假设残差服从正态分布 $N(0, \sigma^2)$ ，残差 $\varepsilon^{(i)}$ 对应的概率密度为：

$$f(\varepsilon^{(i)}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \quad (39)$$

似然函数则可以写成：

$$L(\theta_0, \theta_1) = \prod_{i=1}^n \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right) \right\} = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (\varepsilon^{(i)})^2}{2\sigma^2}\right) \quad (40)$$

用矩阵运算表达上式得到：

$$L(\theta_0, \theta_1) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{2\sigma^2}\right) \quad (41)$$

对数似然函数则可以写成：

$$\ln L(\theta_0, \theta_1) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{2\sigma^2} \quad (42)$$

对数似然函数进一步整理为：

$$\ln L(\theta_0, \theta_1) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \quad (43)$$

对数似然函数对 $\boldsymbol{\theta}$ 求导为 $\boldsymbol{\theta}$ 得到等式：

$$\frac{1}{2\sigma^2} (2\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} - 2\mathbf{X}^\top \mathbf{y}) = 0 \quad (44)$$

整理得到：

$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\theta} = \mathbf{X}^\top \mathbf{y} \quad (45)$$

如果 $\mathbf{X}^\top \mathbf{X}$ 可逆，则 $\boldsymbol{\theta}$ 为：

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (46)$$

这和本章前文的优化解一致。



此外，线性回归还可以从最大后验概率估计 MAP 角度理解，这是《数据有道》要介绍的话题之一。



在代数、线性代数、优化、投影、QR 分解、SVD 分解几个视角基础上，这一章又提供理解线性回归两个新视角——条件概率、最大似然估计 MLE。

为了保证线性回归模型的有效性和精度，通常需要满足以下假设条件：a) 线性关系：自变量和因变量之间的关系必须是线性的。b) 独立性：观测值之间必须是独立的。c) 方差齐性：每个自变量的方差大小相近。d) 误差服从正态分布。e) 自变量之间不能有高度相关性或共线性，因为这将导致模型出现多重共线性，从而使得参数估计变得不稳定。

如果这些假设条件得到满足，那么线性回归模型将会给出较为准确和可靠的结果，否则模型的效果可能会受到影响。

“鸢尾花书”有关线性回归的内容并没有完全结束。图 14 所示为某个线性回归结果。给大家留个悬念，本系列丛书《数据有道》一册将讲解如何理解图 14 结果。

此外，《数据有道》将铺开介绍更多回归算法，比如多元回归分析、正则化、岭回归、套索回归、弹性网络回归、贝叶斯回归、多项式回归、逻辑回归，以及基于主成分分析的正交回归、主元回归等算法。

OLS Regression Results						
Dep. Variable:	AAPL	R-squared:	0.687			
Model:	OLS	Adj. R-squared:	0.686			
Method:	Least Squares	F-statistic:	549.7			
Date:	XXXXXXXXXX	Prob (F-statistic):	4.55e-65			
Time:	XXXXXXXXXX	Log-Likelihood:	678.03			
No. Observations:	252	AIC:	-1352.			
Df Residuals:	250	BIC:	-1345.			
Df Model:	1					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	0.0018	0.001	1.759	0.080	-0.000	0.004
SP500	1.1225	0.048	23.446	0.000	1.028	1.217
Omnibus:	52.424	Durbin-Watson:	1.864			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	210.803			
Skew:	0.777	Prob (JB):	1.68e-46			
Kurtosis:	7.203	Cond. No.	46.1			

图 14. 线性回归结果

25

Principal Component Analysis

主成分分析

以概率统计、几何、矩阵分解、优化为视角



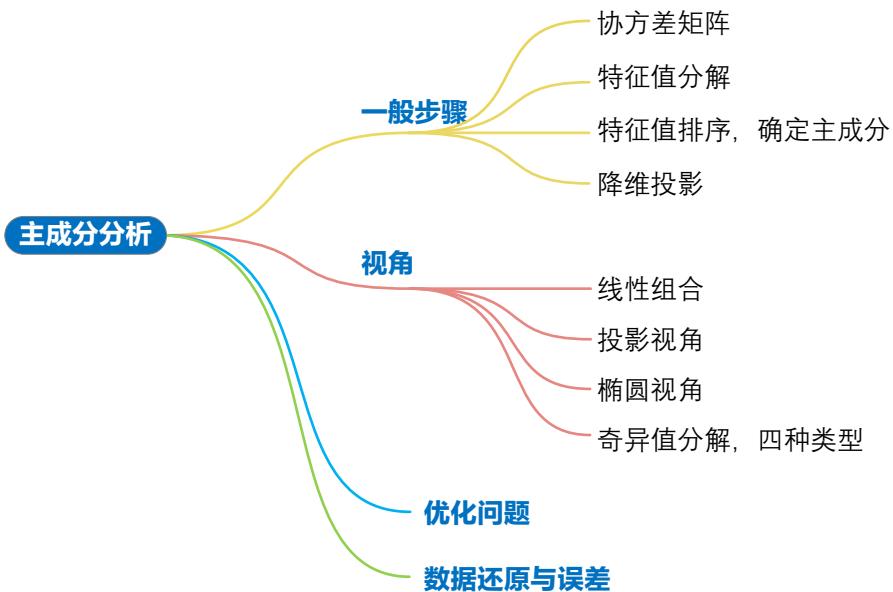
掌握我们的命运的不是星象，而是我们自己。

It is not in the stars to hold our destiny but in ourselves.

—— 威廉·莎士比亚 (William Shakespeare) | 英国剧作家 | 1564 ~ 1616



- ◀ numpy.cov() 计算协方差矩阵
- ◀ numpy.linalg.eig() 特征值分解
- ◀ numpy.linalg.svd() 奇异值分解
- ◀ numpy.random.multivariate_normal() 产生多元正态分布随机数
- ◀ seaborn.heatmap() 绘制热图
- ◀ seaborn.jointplot() 绘制联合分布/散点图和边际分布
- ◀ seaborn.kdeplot() 绘制 KDE 核概率密度估计曲线
- ◀ seaborn.pairplot() 绘制成对分析图
- ◀ sklearn.decomposition.PCA() 主成分分析函数



25.1 再聊主成分分析

主成分分析 (Principal Component Analysis, PCA) 是重要的降维工具。PCA 可以显著减少数据的维数，同时保留数据中对方差贡献最大的成分。简单来说，PCA 的核心思想是通过线性变换将高维数据映射到低维空间中，使得映射后的数据能够尽可能地保留原始数据的信息，同时去除噪声和冗余信息，从而更好地描述数据的本质特征。

另外，对于多维数据，PCA 可以作为一种数据可视化的工具。



PCA 还可以用来构造回归模型，这是《数据有道》一册要介绍的内容。

本章将以概率统计、几何、矩阵分解、优化为视角给大家全景展示主成分分析。此外，大家可以把这一章看成丛书“数学”板块的一个总结。

无监督学习

主成分分析是重要的**无监督学习** (unsupervised learning) 算法。无监督学习是一种机器学习方法，它处理没有标签或输出值的数据。在无监督学习中，模型只能通过分析输入数据的内部结构、模式和相似性来发现数据的特征，从而自动学习数据的潜在结构和规律。

无监督学习通常用于**聚类** (clustering)、**降维** (dimensionality reduction)、**异常检测** (outlier detection) 和**关联规则挖掘** (association rule learning) 等问题。

在聚类问题中，目标是将相似的数据点分组到不同的簇中，从而将数据分割为具有内在结构的不同子集。

在降维问题中，目标是从高维数据中提取出具有代表性的低维特征，从而减少计算复杂度、提高数据可视化效果和去除噪声。主成分分析就是常用的降维算法。

在异常检测问题中，目标是检测数据集中的异常数据点，这些数据点与其它数据点存在显著的差异。本书第 23 章介绍的马氏距离就常用来发现数据中的离群值。

在关联规则挖掘问题中，目标是在大规模数据集中寻找频繁出现的关联项集，从而发现数据中的相关性和关联性。



《数据有道》一册将介绍异常检测、降维、关联规则挖掘等话题，而《机器学习》将关注常见的聚类算法。

一般步骤

如图 1 所示，PCA 的一般步骤如下：

- ◀ 计算原始数据 $X_{n \times D}$ 的协方差矩阵 $\Sigma_{D \times D}$;
- ◀ 对 Σ 特征值分解，获得特征值 λ_i 与特征向量矩阵 $V_{D \times D}$;

- ◀ 对特征值 λ_i 从大到小排序，选择其中特征值最大的 p 个特征向量；
- ◀ 将原始数据（中心化数据）投影到这 p 个正交向量构建的低维空间中，获得得分 $Z_{n \times p}$ 。

很多时候，在第一步中，我们先**标准化** (standardization) 原始数据，即计算 X 的 Z 分数。标准化防止不同特征上方差差异过大。而有些情况，对原始数据 $X_{n \times D}$ 进行中心化 (去均值) 就足够了，即将数据质心移到原点。

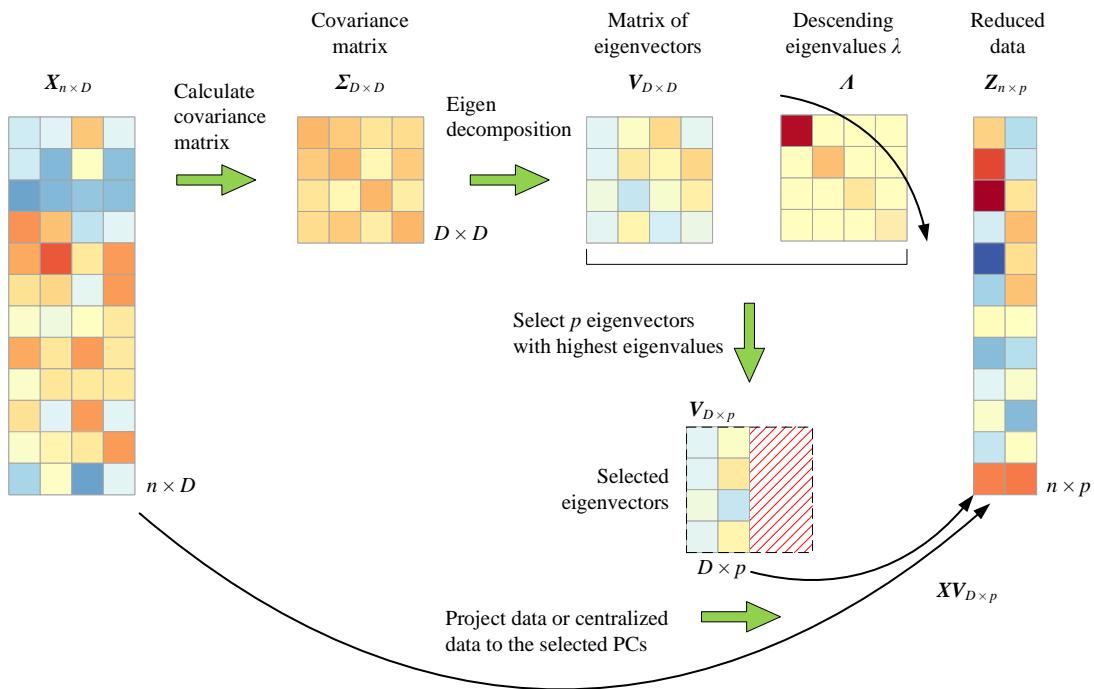


图 1. 主成分分析一般技术路线：特征值分解协方差矩阵



我们在《矩阵力量》第 25 章看到的就是利用标准化数据进行 PCA 分析的技术路线。标准化数据的协方差矩阵实际上就是原数据的相关性系数矩阵。

图 1 所示为通过分解协方差矩阵进行主成分分析过程；当然，也可以通过奇异值分解中心化数据 X_c 进行主成分分析。

25.2 原始数据

《矩阵力量》介绍过，样本数据矩阵 X 可以分别通过行和列来解释。矩阵 X 每一列代表一个特征向量：

$$X = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \mathbf{x}_3 \quad \mathbf{x}_4] \quad (1)$$

X 矩阵每一行代表一个样本。比如， X 矩阵第一行对应是第一个数据点，写成一个行向量 $\mathbf{x}^{(1)}$:

$$\mathbf{x}^{(1)} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \end{bmatrix} \quad (2)$$

图 2 展示原始数据矩阵 X 热图，红色色系代表正数，蓝色色系代表负数，黄色接近 0。 X 矩阵有 12 行，即 12 个样本； X 矩阵有 4 列，即 4 个特征。

⚠ 注意，本例中假设 X 已经中心化 $E(X) = \mathbf{0}^T$ ，即质心位于原点。

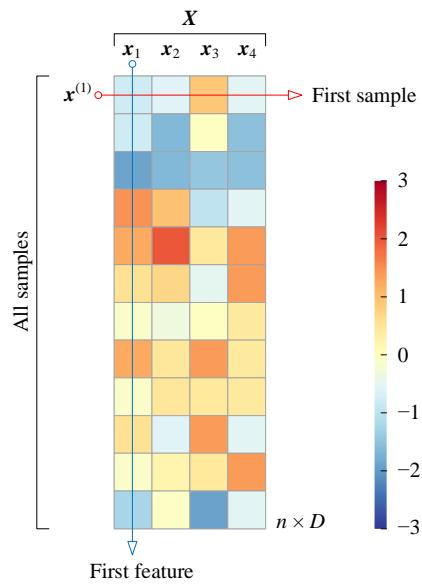
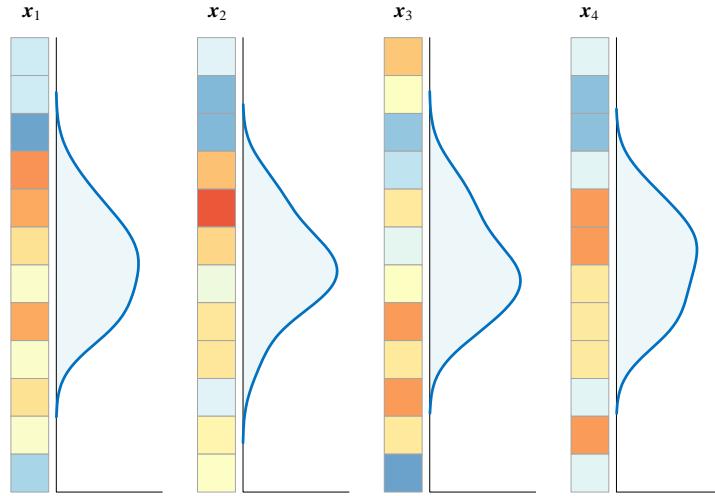


图 2. 原始数据 X 热图， $D = 4$, $n = 12$, X 已经去均值

分布特征

图 3 所示为矩阵 X 每一列特征数据的分布情况；可以发现它们之间的标准差区别不大。但是经过主成分分解之后，大家可以明显发现每一列新特征数据标准差大小差异明显。

图 3. X 四个特征向量数据分布

25.3 特征值分解协方差矩阵

本书第 13 章介绍过， X 的协方差矩阵 Σ 可以通过下式计算得到：

$$\Sigma = \frac{(X - E(X))^T (X - E(X))}{n-1} = \frac{X_c^T X_c}{n-1} \quad (3)$$

其中， $E(X)$ 也常被称作原始数据 X 的质心； $X - E(X)$ 相当于数据中心化。当 n 足够大，(3) 的分母可以用 n 替换。本例设定 $E(X) = \mathbf{0}^T$ ，即 $X = X_c$ 。

如图 5 所示， Σ 为实数对称矩阵，它的特征值分解（谱分解）可以写作：

$$\Sigma = V \Lambda V^T \quad (4)$$

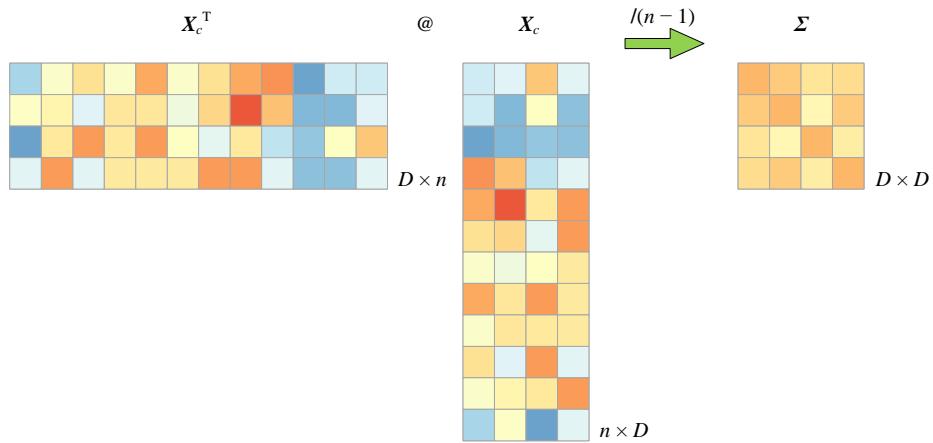
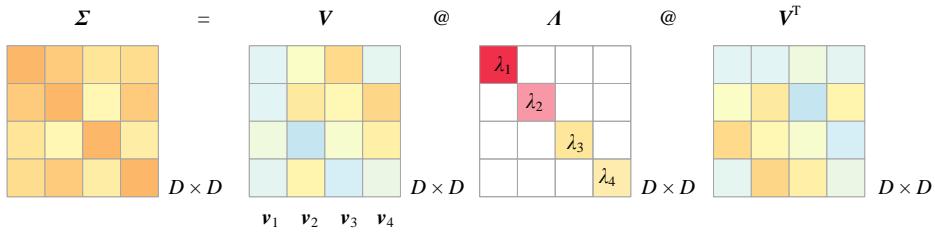
V 为正交矩阵。 V 和自己转置 V^T 乘积为单位阵 I ，即：

$$V^T V = I \quad (5)$$

特征值方阵 Λ 主对角线元素为特征值 λ ，特征值从大到小排列：

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{bmatrix}, \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \quad (6)$$

本书前文介绍过，从统计学角度来讲， λ_j 是第 j 个主成分所贡献的方差。

图 4. 计算原始数据协方差矩阵, $D = 4$, $n = 12$ 图 5. 协方差矩阵特征值分解, $D = 4$

主成分、载荷

V 为特征向量构造的 $D \times D$ 的方阵:

$$V = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_D \end{bmatrix}_{\text{PC1 PC2}} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,D} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ v_{D,1} & v_{D,2} & \cdots & v_{D,D} \end{bmatrix} \quad (7)$$

\mathbf{v}_1 被称作**第一主成分** (first principal component), 本书常记做 PC1; \mathbf{v}_2 被称作**第二主成分** (second principal component), 记做 PC2; 以此类推。

V 的列向量也叫**载荷** (loadings)。注意, 有些文献中载荷定义为:

$$V\sqrt{\Lambda} = [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_D] \begin{bmatrix} \sqrt{\lambda_1} & & & \\ & \sqrt{\lambda_2} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_D} \end{bmatrix} = [\sqrt{\lambda_1}\mathbf{v}_1 \quad \sqrt{\lambda_2}\mathbf{v}_2 \quad \cdots \quad \sqrt{\lambda_D}\mathbf{v}_D] \quad (8)$$

迹, 总方差

本书前文介绍过，协方差矩阵 Σ 的迹 $\text{trace}(\Sigma)$ 等于的特征值方阵 A 迹 $\text{trace}(A)$ ：

$$\text{trace}(\Sigma) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_D^2 = \sum_{j=1}^D \sigma_j^2 = \text{trace}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_D = \sum_{j=1}^D \lambda_j \quad (9)$$

第 j 个特征值 λ_j 对**方差总和** (total variance) 的贡献百分比为：

$$\frac{\lambda_j}{\sum_{i=1}^D \lambda_i} \times 100\% \quad (10)$$

前 p 个特征值，即 p 个主成分**总方差解释** (total variance explained) 的百分比为：

$$\frac{\sum_{j=1}^p \lambda_j}{\sum_{i=1}^D \lambda_i} \times 100\% \quad (11)$$

"total variance" 指的是原始数据中所有变量的总方差，"explained" 意味着这个方差被 PCA 模型中所选的主成分所解释。因此，"total variance explained" 表示通过 PCA 转换后的主成分所解释的原始数据中总方差的比例。这个值通常以百分比的形式给出，可以帮助我们了解每个主成分对数据的解释程度，以及所有主成分的总体效果。

主成分分析中，我们常用**陡坡图** (scree plot) 可视化这个百分比。



《数据有道》一册中大家会看到很多陡坡图实例。

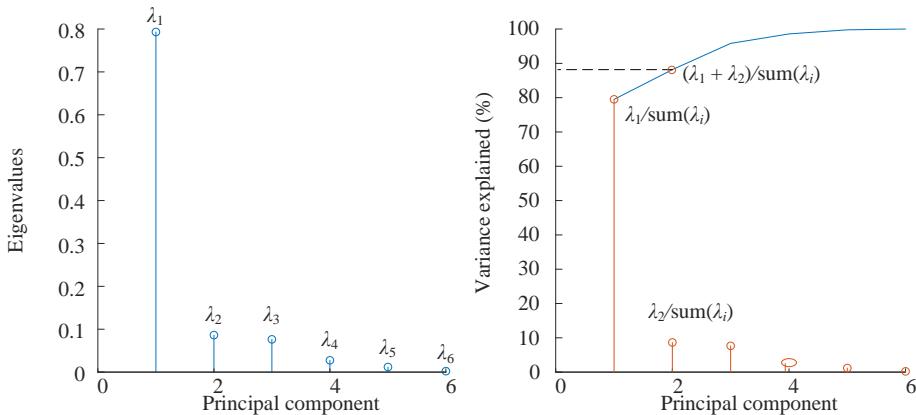


图 6. PCA 分析主元方差和陡坡图

25.4 投影

本节从投影角度介绍 PCA。数据矩阵 X 投影到矩阵 V 正交系 (v_1, v_2, \dots, v_D) 得到新特征数据矩阵 Z ，即：

$$\mathbf{Z} = \mathbf{X}\mathbf{V} \quad (12)$$

\mathbf{V} 常被称作**载荷** (loadings), \mathbf{Z} 常被称作**得分** (scores)。图 7 所示 $\mathbf{Z} = \mathbf{X}\mathbf{V}$ 矩阵运算原理图。

→ 《矩阵力量》第 10 章特别介绍过这种数据投影，建议大家回顾。

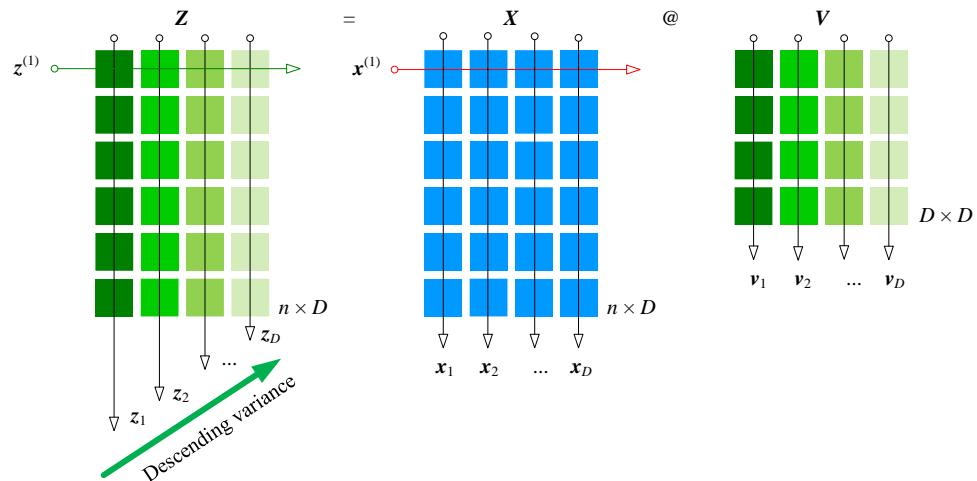


图 7. PCA 分解数据关系 $\mathbf{Z} = \mathbf{X}\mathbf{V}$

图 8 所示为将图 2 给出数据矩阵 \mathbf{X} 投影到矩阵 \mathbf{V} ，得到的得分 \mathbf{Z} 。

⚠ 值得强调的一点是，把原始数据 \mathbf{X} 或中心化数据 \mathbf{X}_c 投影到 \mathbf{V} 中结果不一样。从统计角度来看，差异主要体现在质心位置，而投影得到的数据协方差矩阵相同。

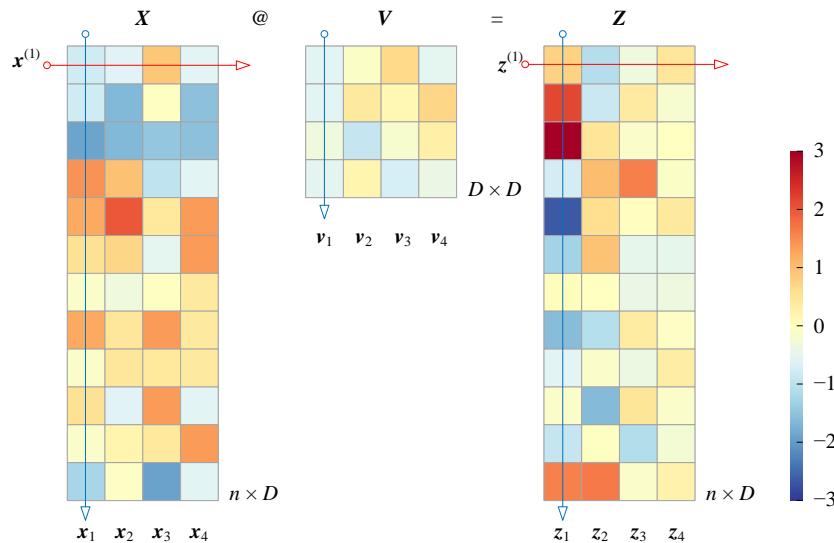


图 8. \mathbf{Z} 、 \mathbf{X} 和 \mathbf{V} 这三个矩阵关系和热图

Z的列向量

前文讨论过，矩阵 X 每一列特征数据方差区别不大（见图 3）；而图 9 告诉我们，经过 PCA 分解得到的矩阵 Z 四个新特征数据分布差异显著。

如图 9 所示，第一列 z_1 数据分布最为分散，也就是**第一主成分** (first principal component) 解释了数据中最多方差。第一列 z_1 到第四列 z_4 数据分散情况逐渐降低，热图对应的色差从明显到模糊。

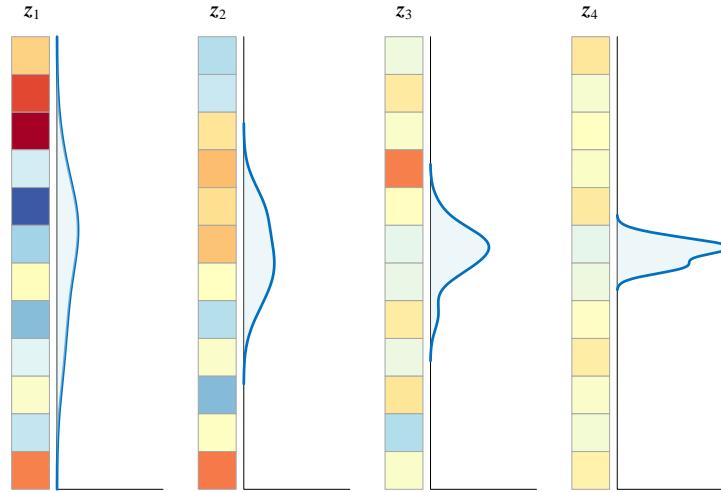


图 9. Z 四个新特征数据分布

将 (12) 展开得到：

$$[z_1 \ z_2 \ \cdots \ z_D] = X \begin{bmatrix} v_1 & v_2 & \cdots & v_D \\ \text{PC1} & \text{PC2} & & \end{bmatrix} \quad (13)$$

由此，得到图 10 所示主成分分析运算的数据关系：

$$\begin{cases} z_1 = Xv_1 \\ z_2 = Xv_2 \\ \vdots \\ z_D = Xv_D \end{cases} \quad (14)$$

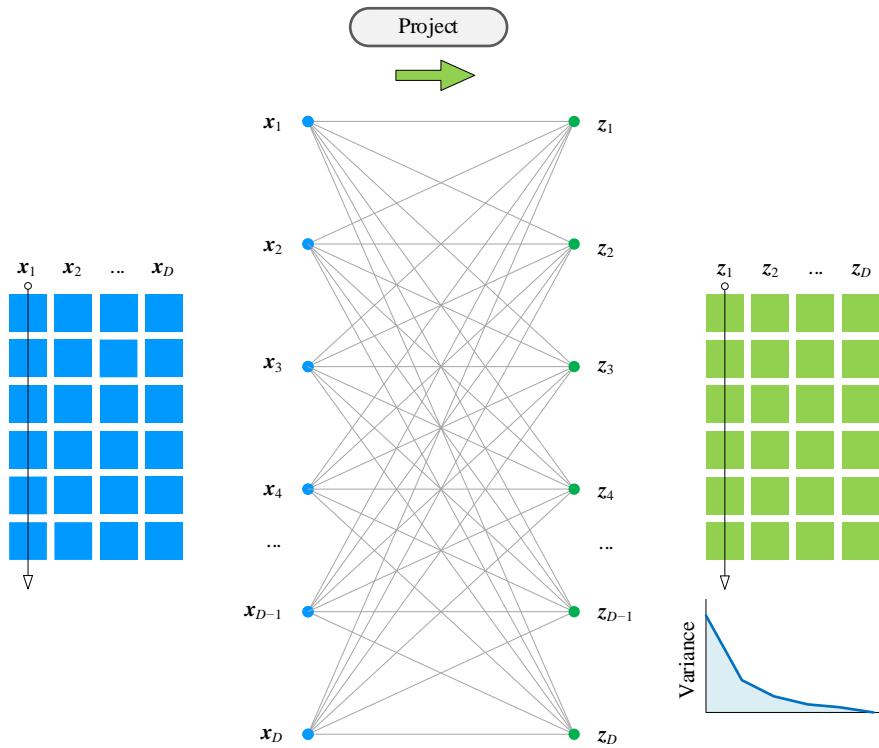


图 10. PCA 中数据关系

线性组合

如图 11 所示，以列向量 v_1 为例，它的每个元素相当于 $[x_1, x_2, \dots, x_D]$ 线性组合对应系数。将 X 向 v_1 投影：

$$z_1 = Xv_1 \quad (15)$$

(15) 展开得到：

$$z_1 = [x_1 \ x_2 \ \cdots \ x_D] \begin{bmatrix} v_{1,1} \\ v_{2,1} \\ \vdots \\ v_{D,1} \end{bmatrix} = v_{1,1}x_1 + v_{2,1}x_2 + \cdots + v_{D,1}x_D \quad (16)$$

$v_1, \text{PC1}$

简单来讲， z_1 相当于 $[x_1, x_2, \dots, x_D]$ 的某种特殊线性组合。

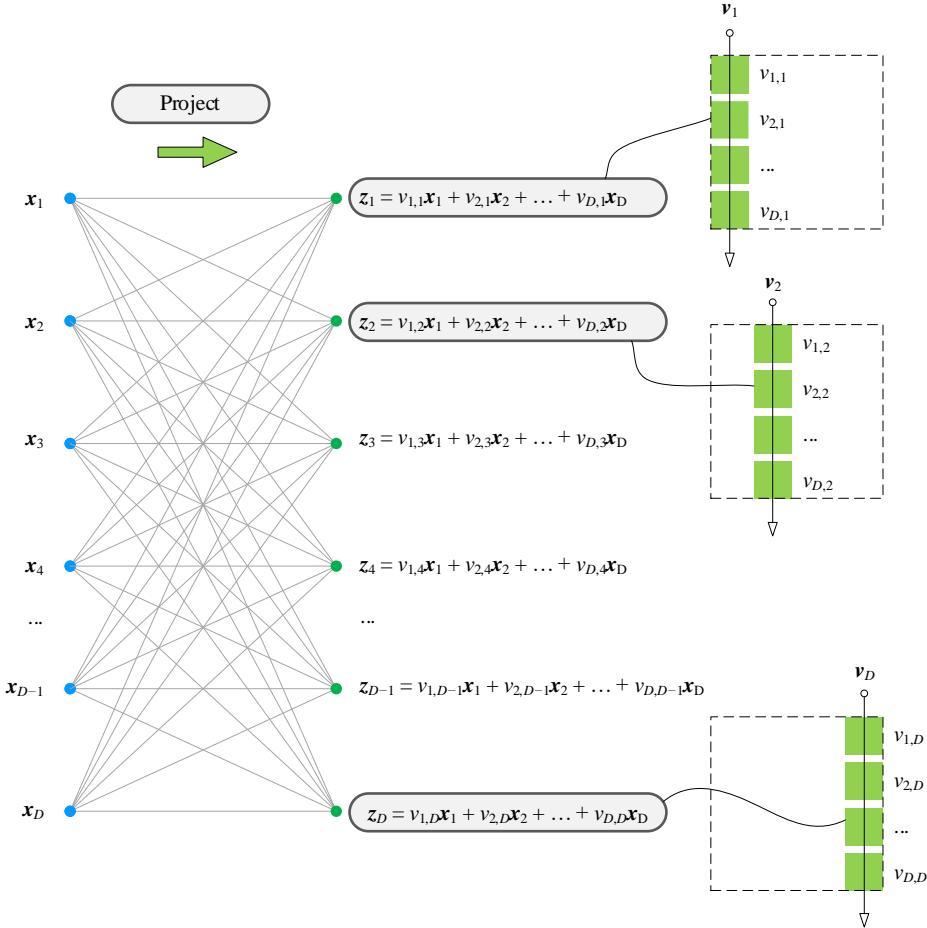
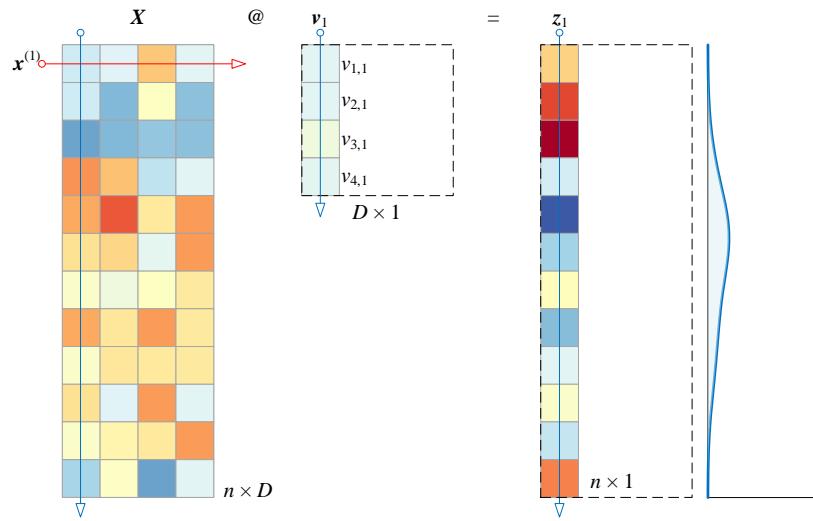
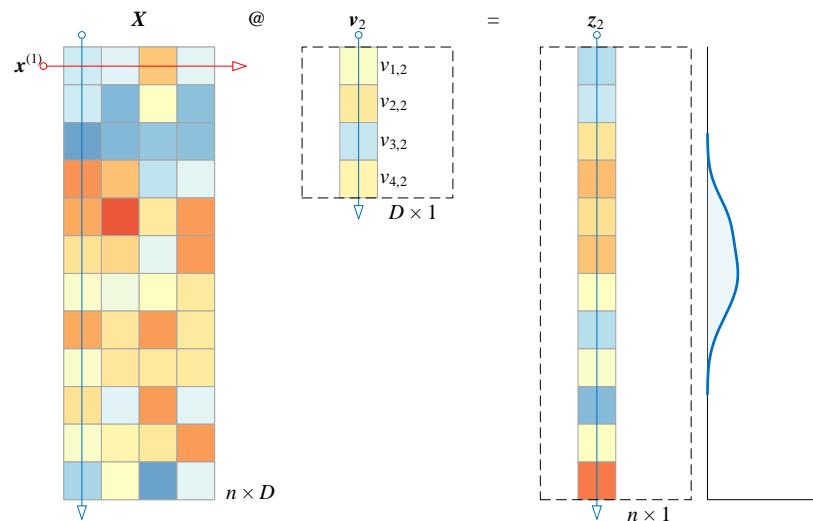


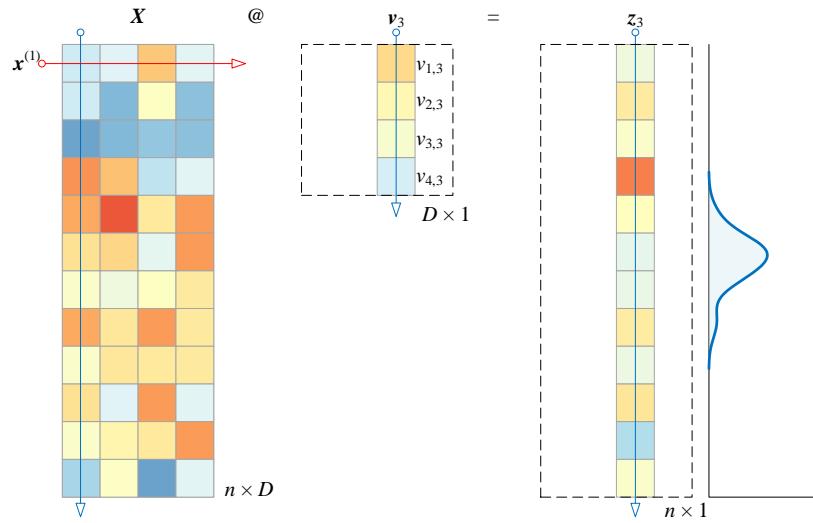
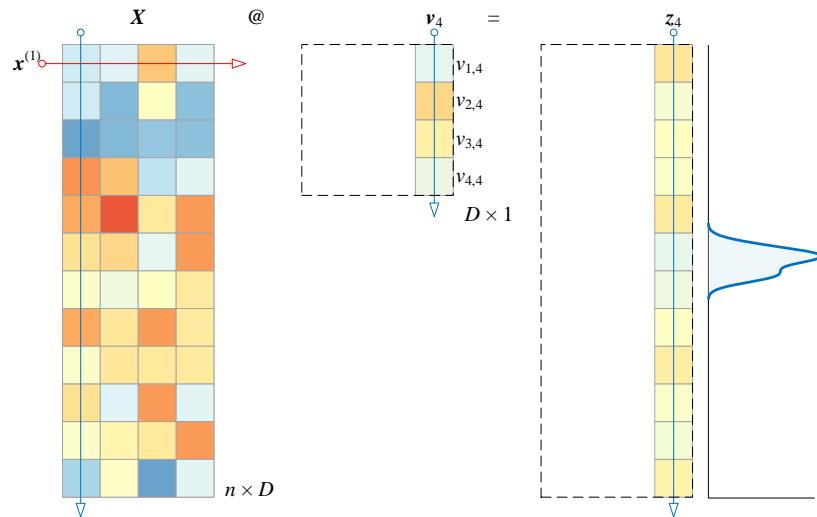
图 11. 线性组合角度看 PCA

朝向量投影

图 12 ~ 图 15 分别展示数据矩阵 X 向 v_1, v_2, v_3 和 v_4 向量投影。

图 12 所示 $z_1 = Xv_1$ 运算相当于数据 X 向 v_1 向量 (第一主成分) 投影获得 z_1 。图 13 展示 $z_2 = Xv_2$ 运算等价于数据 X 向 v_2 (第二主成分) 投影获得 z_2 。以此类推。

图 12. 数据 X 向 v_1 向量投影图 13. 数据 X 向 v_2 向量投影

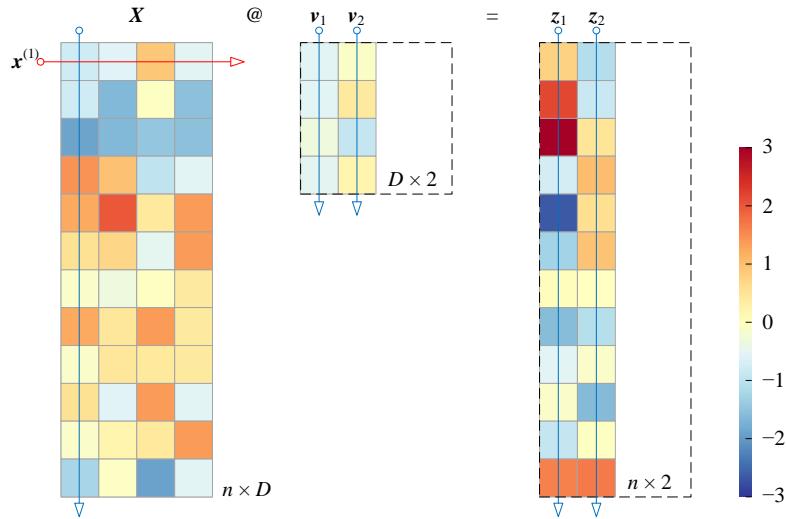
图 14. 数据 X 向 v_3 向量投影图 15. 数据 X 向 v_4 向量投影

朝平面投影

同样， $[z_1, z_2]$ 是 X 向 $[v_1, v_2]$ 投影结果，即四维数据 X 向二维空间投影。运算过程如下：

$$[z_1 \ z_2] = X [v_1 \ v_2] \quad (17)$$

图 16 所示为 (17) 运算过程及结果热图。

图 16. 数据 X 向 $[v_1, v_2]$ 投影

Z 的协方差矩阵

前文假设 X 已经中心化，因此 z_1 的期望值为 0。对 z_1 求方差，可以得到：

$$\text{var}(z_1) = \frac{(Xv_1)^T(Xv_1)}{n-1} = \frac{v_1^T X^T X v_1}{n-1} = v_1^T \frac{X^T X}{n-1} v_1 = v_1^T \Sigma v_1 \quad (18)$$

类似地，

$$\text{var}(z_2) = v_2^T \Sigma v_2, \dots, \text{var}(z_D) = v_D^T \Sigma v_D \quad (19)$$

这样， Z 的协方差矩阵可以通过下式计算得到：

$$\begin{aligned} \text{var}(Z) &= \frac{(XV)^T(XV)}{n-1} = \frac{V^T X^T X V}{n-1} \\ &= V^T \frac{X^T X}{n-1} V = V^T \Sigma V = \begin{bmatrix} v_1^T \Sigma v_1 & & & \\ & v_2^T \Sigma v_2 & & \\ & & \ddots & \\ & & & v_D^T \Sigma v_D \end{bmatrix} = A = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix} \end{aligned} \quad (20)$$

观察 (20) 所示协方差矩阵，可以发现主对角线以外元素均为 0，也就是 Z 的列向量两两正交（前提是其质心位于原点），线性相关系数为 0。

$Z_{n \times p}$ 的协方差矩阵为：

$$\text{var}(Z_{n \times p}) = \frac{(XV_{D \times p})^T(XV_{D \times p})}{n-1} = V_{D \times p}^T \frac{X^T X}{n-1} V_{D \times p} = V_{D \times p}^T \Sigma V_{D \times p} = A_{p \times p} = \begin{bmatrix} \lambda_1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix} \quad (21)$$

→ 对于投影数据的方差计算，我们已经在本书第 14 章详细介绍过，记忆模糊的话请自行回顾复习。

25.5 几何视角看 PCA

如图 17 所示，椭圆中心对应质心 μ ，椭圆和 $\pm\sigma$ 标准差构成的矩形相切，四个切点分别为 A、B、C 和 D，对角切点两两相连得到两条直线 AC、BD。

本书前文介绍过，AC 相当于在给定 X_2 条件下 X_1 的条件概率期望值；BD 相当于在给定 X_1 条件下 X_2 的条件概率期望值。

图 17 中，EF 为椭圆长轴；FH 为椭圆短轴。而 EF 就相当于 PCA 的第一主成分，FH 为第二主成分。

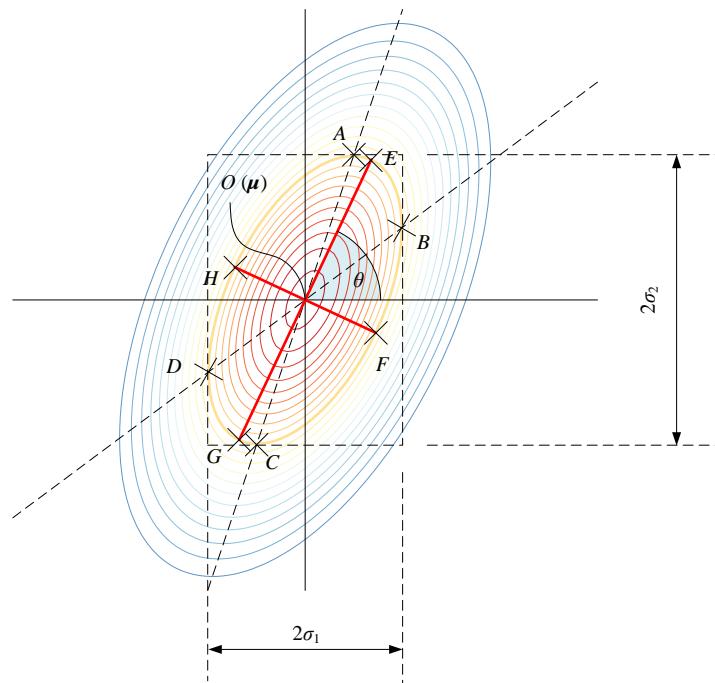


图 17. 主成分分析和椭圆的关系

图 18 则从椭圆视角解释主成分分析。假设图 18 原始数据已经标准化，计算得到协方差矩阵 Σ ，找到 Σ 对应椭圆的半长轴所在方向 v_1 。 v_1 对应的便是第一主成分 PC1。原始数据朝 v_1 投影得到的数据对应最大方差。

→ 整个过程实际上用到了“鸢尾花书”《矩阵力量》一本中介绍的平移、缩放、正交化、投影、旋转等数学工具。

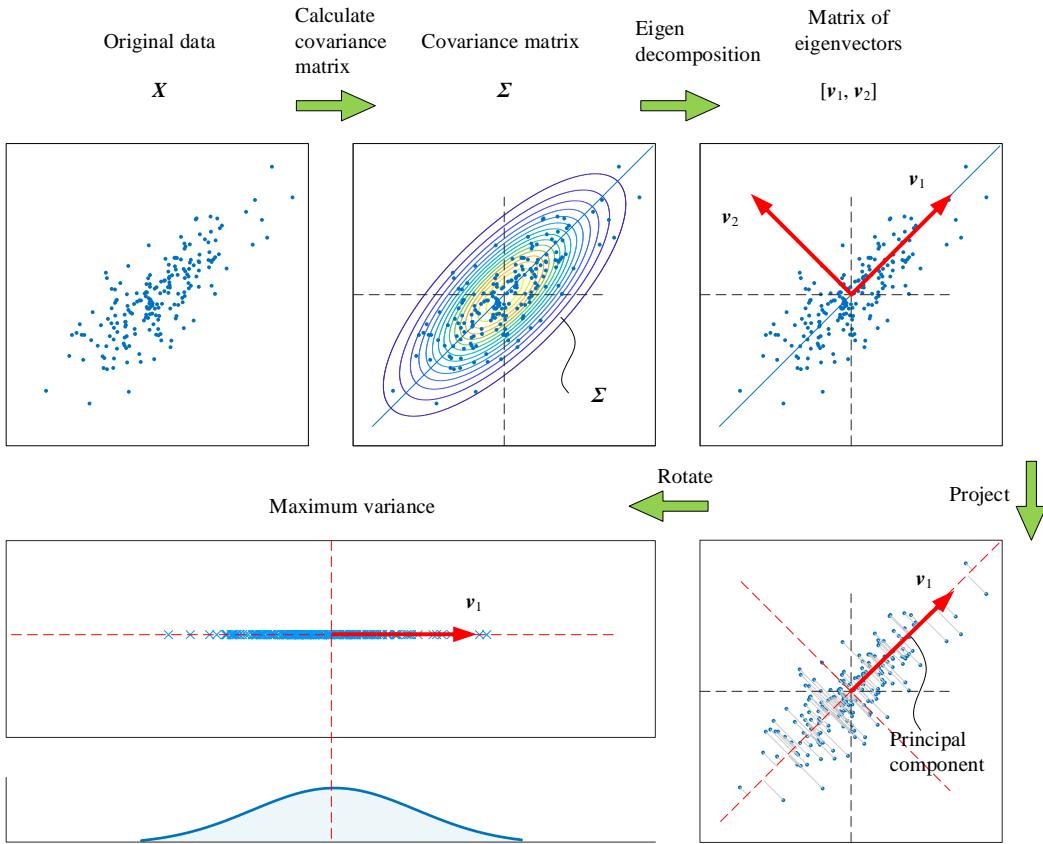


图 18. 几何视角下通过特征值分解协方差矩阵进行主成分分析

如图 19 所示，从线性变换角度来看，主成分分析无非就是在不同的坐标系中看同一组数据。数据朝不同方向投影会得到不同的投影结果，对应不同的分布；朝椭圆长轴方向投影，得到的数据标准差最大；朝椭圆短轴方向投影得到的数据标准差最小。

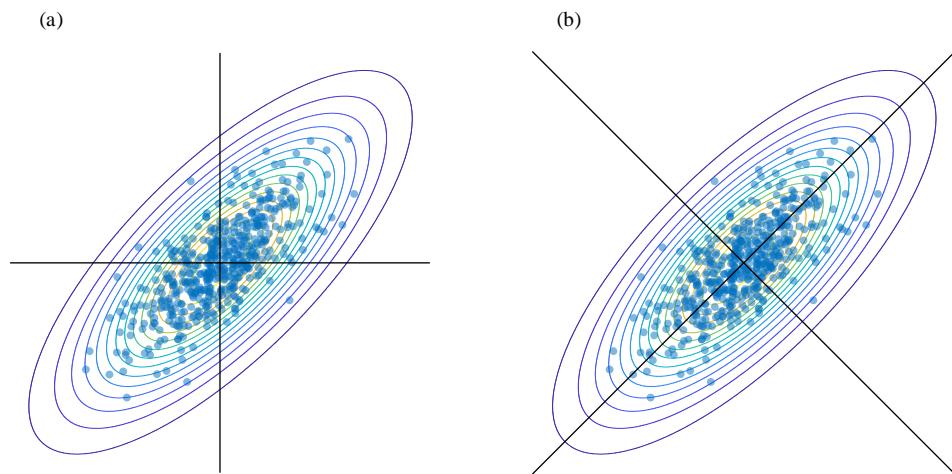


图 19. 两个角度看数据

举个例子

图 20 (a) 所示为原始二维数据 X 的散点图，可以发现数据的质心位于 $[1, 2]^T$ 。分析数据 X ，可以发现数据的两个特征上分布分散情况相似，也就是方差大小几乎相同。

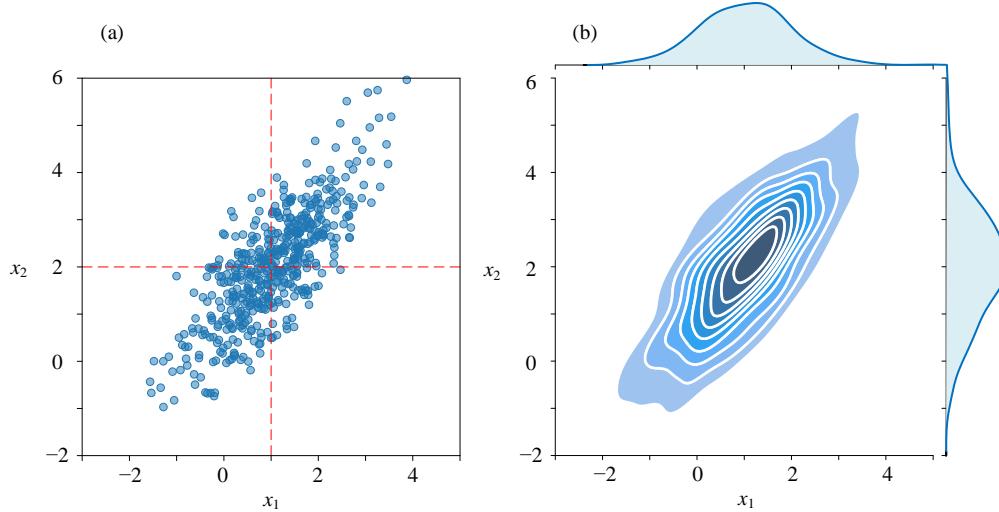


图 20. 原始二维数据 X

利用 `sklearn.decomposition.PCA()` 函数，我们可以通过 `pca.components_` 获得主成分向量。利用 `pca.transform(X)` 可以获得投影后的数据 Y 。图 21 对比 Y 两列数据分布。图 22 所示为数据 Y 在 $[v_1, v_2]$ 中散点图。

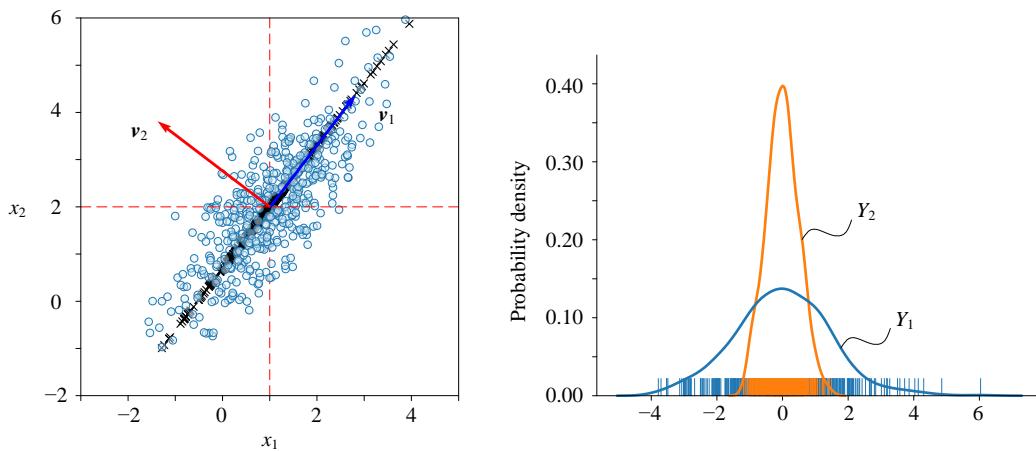
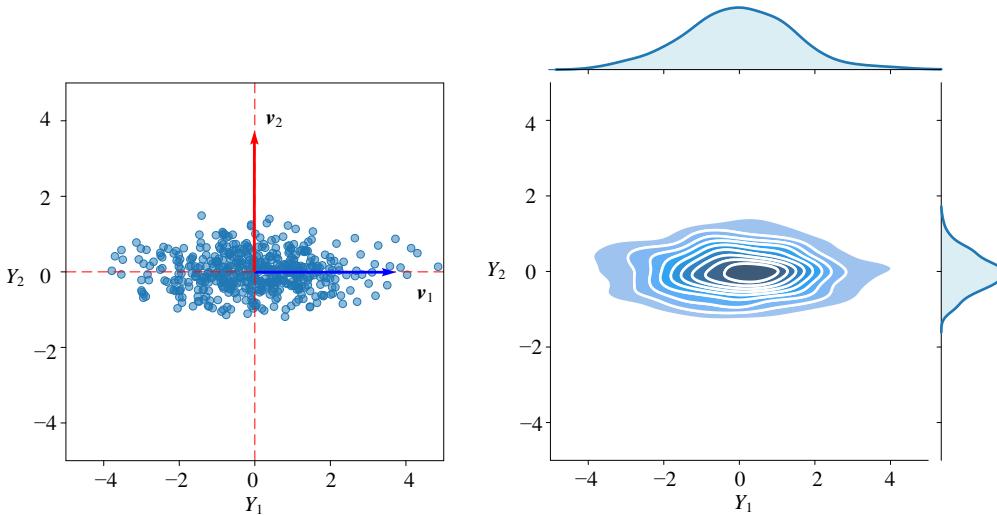
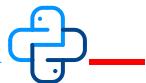


图 21. 主成分数据分布

图 22. 数据 \mathbf{Y} 在 $[v_1, v_2]$ 中散点图

Bk5_Ch25_01.py 绘制图 20 ~ 图 22。

25.6 奇异值分解

四种奇异值分解

奇异值分解 (singular value decomposition, SVD) 也可以用来做主成分分析。丛书在《矩阵力量》一本系统讲解过奇异值分解的四种类型：

- ◀ **完全型** (full);
- ◀ **经济型** (economy-size, thin);
- ◀ **紧凑型** (compact);
- ◀ **截断型** (truncated)。

如图 23 所示，完全型奇异值分解中， \mathbf{U} 为方阵， \mathbf{S} 矩阵并非方阵。

$$\begin{array}{c}
 X = U @ S @ V^T \\
 \begin{array}{c|c|c|c}
 \text{---} & \text{---} & \text{---} & \text{---} \\
 \end{array}
 \end{array}$$

X $=$ U $@$ S $@$ V^T
 $n \times D$ $n \times n$ $n \times n$ $D \times D$

图 23. 完全 (full) 奇异值分解

去掉图 23 中这个全 0 矩阵 O , 便得到经济型奇异值分解, 具体如图 24 所示。经济型 SVD 中, U 的形状和 X 相同, S 矩阵为对角方阵, 形状为 $D \times D$ 。

$$\begin{array}{c}
 X = U @ S @ V^T \\
 \begin{array}{c|c|c|c}
 \text{---} & \text{---} & \text{---} & \text{---} \\
 \end{array}
 \end{array}$$

X $=$ U $@$ S $@$ V^T
 $n \times D$ $n \times D$ $D \times D$ $D \times D$

图 24. 经济型奇异值分解

当 X 非满秩时, 即 $\text{rank}(X) = r < D$, 图 24 经济型奇异值分解可以进一步简化为如图 25 所示的紧凑型 SVD 分解。

$$\begin{array}{c}
 X = U_{n \times r} @ S_{r \times r} @ V_{r \times D}^T \\
 \begin{array}{c|c|c|c}
 \text{---} & \text{---} & \text{---} & \text{---} \\
 \end{array}
 \end{array}$$

X $=$ $U_{n \times r}$ $@$ $S_{r \times r}$ $@$ $V_{r \times D}^T$
 $n \times D$ $n \times r$ $r \times r$ $r \times D$

图 25. 紧凑型奇异值分解, X 非满秩

在线性代数中，矩阵的秩指的是其列向量或行向量的线性无关的数量。如果矩阵的秩等于它的行数或列数中的较小值，则称该矩阵为满秩矩阵。如果矩阵的秩小于它的行数或列数中的较小值，则称该矩阵为非满秩矩阵。

在机器学习中，非满秩的矩阵通常表示存在冗余或线性相关的特征或样本。这些冗余或线性相关的特征或样本可能会导致算法的过拟合，降低模型的准确性和稳定性。因此，在许多机器学习算法中，对于非满秩矩阵，通常需要进行一些特殊的处理，例如降维或正则化，以减少冗余或相关性，并提高模型的效果。

图 26 给出的是截断型奇异值分解， $S_{p \times p}$ 仅使用图 24 中 S 矩阵 p 个主成分特征值，形状为 $p \times p$ 。注意，图 26 中使用的是约等号“ \approx ”；这是因为，约等号右侧矩阵运算仅仅还原 X 矩阵部分数据，并非还原全部信息。本章后续将会展开讲解数据还原和误差。

图 26. 截断型奇异值分解

SVD 完成主成分分析

首先中心化(去均值)数据矩阵。对已经去均值的矩阵 $X_{n \times D}$ 进行完全型 SVD 分解，得到：

$$X = USV^T \quad (22)$$

V 和 U 均为正交矩阵，即满足：

$$\begin{aligned} UU^T &= U^T U = I \\ VV^T &= V^T V = I \end{aligned} \quad (23)$$

Python 中常用奇异值分解函数为 `numpy.linalg.svd()`。

由于 X 已经中心化，其协方差矩阵可以通过下式计算获得：

$$\Sigma = \frac{X^T X}{n-1} \quad (24)$$

将(22) 代入 (24) 得到：

$$\Sigma = \frac{(\mathbf{U}\mathbf{S}\mathbf{V}^T)^T \mathbf{U}\mathbf{S}\mathbf{V}^T}{n-1} = \frac{\mathbf{V}\mathbf{S}^T\mathbf{S}\mathbf{V}^T}{n-1} \quad (25)$$

对协方差矩阵进行特征值分解：

$$\Sigma = \mathbf{V}\mathbf{A}\mathbf{V}^T \quad (26)$$

联立 (25) 和 (26)，

$$\frac{\mathbf{V}\mathbf{S}^T\mathbf{S}\mathbf{V}^T}{n-1} = \mathbf{V}\mathbf{A}\mathbf{V}^T \quad (27)$$

对于经济型 SVD 分解， \mathbf{S} 为对角方阵，(27) 整理得到：

$$\frac{\mathbf{S}^2}{n-1} = \mathbf{A} \quad (28)$$

即

$$\frac{1}{n-1} \begin{bmatrix} s_1^2 & & & \\ & s_2^2 & & \\ & & \ddots & \\ & & & s_D^2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_D \end{bmatrix} \quad (29)$$

注意， $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$ 。

奇异值和特征值存在如下关系：

$$\frac{s_j^2}{n-1} = \lambda_j \quad (30)$$

s_j 为第 j 个主成分的**奇异值** (singular value)， λ_j 为协方差矩阵的第 j 个特征值。

理解 \mathbf{U}

\mathbf{Z} 可以还原 \mathbf{X} ：

$$\mathbf{X} = \mathbf{Z}\mathbf{V}^{-1} = \mathbf{Z}\mathbf{V}^T \quad (31)$$

对比 (22) 和 $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ ，可以发现：

$$\mathbf{Z} = \mathbf{U}\mathbf{S} \quad (32)$$

也就是

$$[\mathbf{z}_1 \ \mathbf{z}_2 \ \dots \ \mathbf{z}_D] = [\mathbf{u}_1 \ \mathbf{u}_2 \ \dots \ \mathbf{u}_D] \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix} = [s_1 \mathbf{u}_1 \ s_2 \mathbf{u}_2 \ \dots \ s_D \mathbf{u}_D] \quad (33)$$

即：

$$s_1 \mathbf{u}_1 = \mathbf{z}_1, \quad s_2 \mathbf{u}_2 = \mathbf{z}_2, \dots \quad (34)$$

对 \mathbf{z}_1 求方差：

$$\text{var}(\mathbf{z}_1) = \frac{\mathbf{z}_1^T \mathbf{z}_1}{n-1} = \frac{(\mathbf{s}_1 \mathbf{u}_1)^T (\mathbf{s}_1 \mathbf{u}_1)}{n-1} = \frac{s_1^2 \|\mathbf{u}_1\|^2}{n-1} = \frac{s_1^2}{n-1} = \lambda_1 \quad (35)$$

可以发现矩阵 \mathbf{U} 每一列数据相当于 \mathbf{Z} 对应列向量的标准化：

$$\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_D] = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_D \\ s_1 & s_2 & \cdots & s_D \end{bmatrix} \quad (36)$$

也就是：

$$\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_D] = \mathbf{Z} \mathbf{S}^{-1} \quad (37)$$

至此，我们理解了 SVD 分解中矩阵 \mathbf{U} 的内涵。

张量积

用张量积来展开 SVD 分解：

$$\begin{aligned} \mathbf{X} &= \mathbf{U} \mathbf{S} \mathbf{V}^T \\ &= [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_D] \begin{bmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_D \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix} \\ &= s_1 \mathbf{u}_1 \mathbf{v}_1^T + s_2 \mathbf{u}_2 \mathbf{v}_2^T + \cdots + s_D \mathbf{u}_D \mathbf{v}_D^T \\ &= s_1 \mathbf{u}_1 \otimes \mathbf{v}_1 + s_2 \mathbf{u}_2 \otimes \mathbf{v}_2 + \cdots + s_D \mathbf{u}_D \otimes \mathbf{v}_D \end{aligned} \quad (38)$$

图 27 所示为 (38) 还原原始数据的过程。

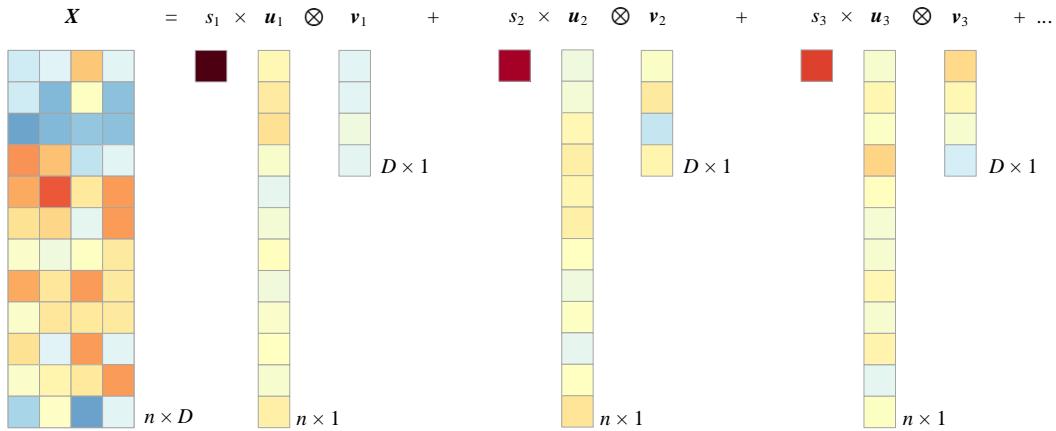


图 27. 张量积 $s_1 \mathbf{u}_1 \otimes \mathbf{v}_1$ 、 $s_2 \mathbf{u}_2 \otimes \mathbf{v}_2$ 等之和还原数据 X

25.7 优化问题

下面我们从优化角度理解 PCA。如图 28 所示， X 为数据矩阵，即 X 质心零向量。 v 为单位向量。数据 X 在 v 上投影结果为 z ，即 $z = Xv$ 。

主成分分析中，选取 v 的标准是—— z 方差最大化。这便是构造 PCA 优化问题的第一个角度。

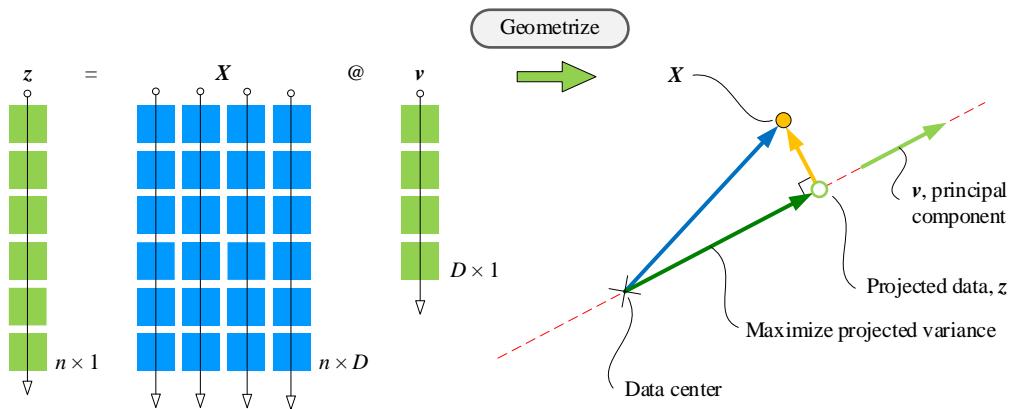


图 28. 主成分分析优化问题

由于 X 为数据矩阵，因此 z 的均值也为 0；因此， z 方差为：

$$\text{var}(z) = \frac{z^T z}{n-1} = v^T \frac{X^T X}{n-1} v \quad (39)$$

发现上式隐藏着数据 X 协方差矩阵，因此 $\text{var}(z)$ 为：

$$\text{var}(z) = v^T \Sigma v \quad (40)$$

v 为单位列向量，即满足如下约束条件：

$$v^T v = 1 \quad (41)$$

有以上分析，我们便可以构造主成分分析优化问题，优化目标为数据在 v 方向上数据投影方差最大化：

$$\begin{aligned} & \arg \max_v v^T \Sigma v \\ & \text{subject to: } v^T v - 1 = 0 \end{aligned} \quad (42)$$

上式最大化优化问题等价于如下最小化优化问题：

$$\begin{aligned} & \arg \min_{\boldsymbol{\nu}} -\boldsymbol{\nu}^T \boldsymbol{\Sigma} \boldsymbol{\nu} \\ & \text{subject to: } \boldsymbol{\nu}^T \boldsymbol{\nu} - 1 = 0 \end{aligned} \quad (43)$$

构造拉格朗日函数 $L(\boldsymbol{\nu}, \lambda)$:

$$L(\boldsymbol{\nu}, \lambda) = -\boldsymbol{\nu}^T \boldsymbol{\Sigma} \boldsymbol{\nu} + \lambda (\boldsymbol{\nu}^T \boldsymbol{\nu} - 1) \quad (44)$$

λ 为拉格朗日乘子。 $L(\boldsymbol{\nu}, \lambda)$ 对 $\boldsymbol{\nu}$ 求偏导，最优解必要条件如下：

$$\nabla_{\boldsymbol{\nu}} L(\boldsymbol{\nu}, \lambda) = \frac{\partial L(\boldsymbol{\nu}, \lambda)}{\partial \boldsymbol{\nu}} = (-2 \boldsymbol{\Sigma} \boldsymbol{\nu} + 2 \lambda \boldsymbol{\nu})^T = \boldsymbol{0} \quad (45)$$

有关拉格朗日乘子法，请大家回顾《矩阵力量》第 18 章。

整理 (45) 得到：

$$\boldsymbol{\Sigma} \boldsymbol{\nu} = \lambda \boldsymbol{\nu} \quad (46)$$

由此， $\boldsymbol{\nu}$ 为数据 X 协方差矩阵 $\boldsymbol{\Sigma}$ 特征向量。 $\text{var}(z)$ 整理为：

$$\text{var}(z) = \boldsymbol{\nu}^T \boldsymbol{\Sigma} \boldsymbol{\nu} = \boldsymbol{\nu}^T \lambda \boldsymbol{\nu} = \lambda \boldsymbol{\nu}^T \boldsymbol{\nu} = \lambda \quad (47)$$

也就是说， $\text{var}(z)$ 最大值对应 $\boldsymbol{\Sigma}$ 最大特征值。这一节从优化角度解释了为什么特征值分解能够完成主成分分析。

25.8 数据还原和误差

还原

前文介绍过， Z 反向可以通过 $X = ZV^T$ 还原 X 。图 29 所示为还原得到 X 过程。图 30 所示热图，矩阵 Z 还原转化为原始数据矩阵 X 。

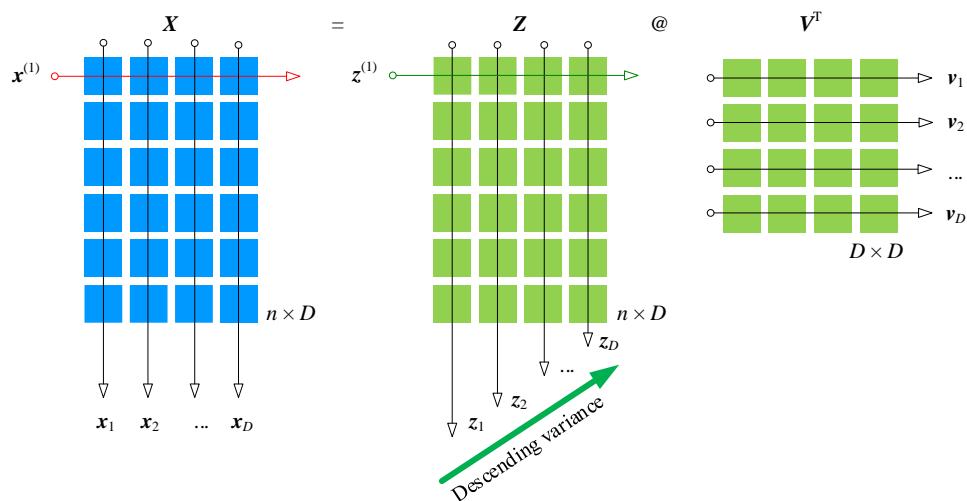
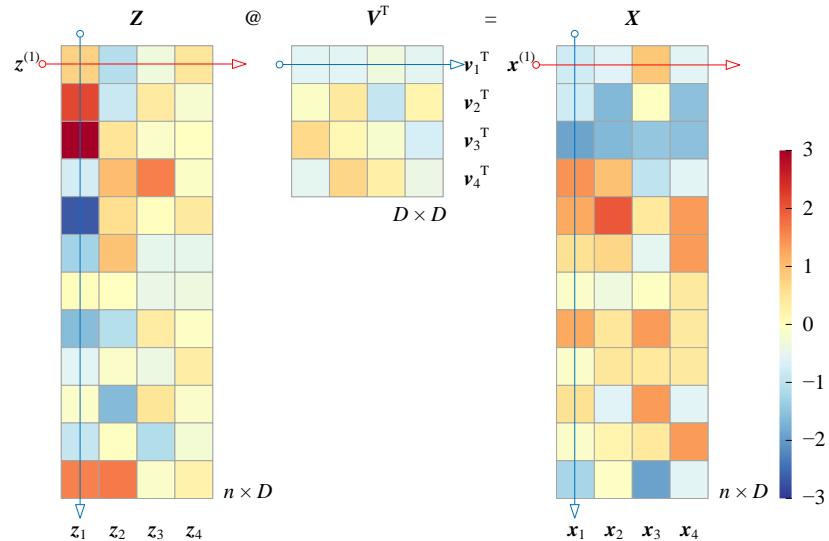


图 29. 反向还原数据 $X = ZV^T$

⚠ 再次强调，图 29 这种还原计算成立的条件是 X 的质心位于原点。

图 30. 新特征数据矩阵 Z 还原转化为原始数据矩阵 X

$X = ZV^T$ 展开得到下式：

$$X = [z_1 \ z_2 \ z_3 \ z_4] \begin{bmatrix} v_1^T \\ v_2^T \\ v_3^T \\ v_4^T \end{bmatrix} = z_1 v_1^T + z_2 v_2^T + z_3 v_3^T + z_4 v_4^T \quad (48)$$

(48) 所示运算过程如图 31 所示。

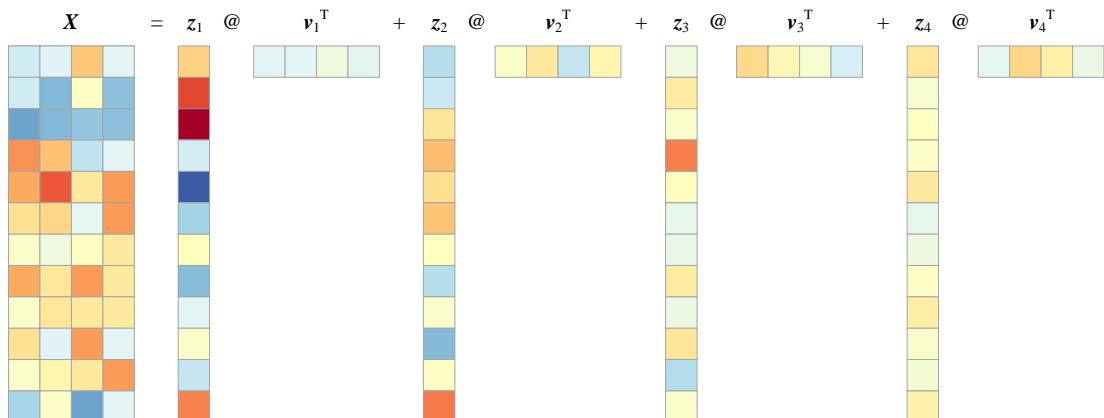


图 31. 还原原始数据运算

图 32 所示为 z_1 还原 X 部分数据，对应运算如下：

$$\mathbf{X}_1 = \mathbf{z}_1 \mathbf{v}_1^T \quad (49)$$

展开上式得到：

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{z}_1 \mathbf{v}_1^T \\ &= \mathbf{z}_1 \begin{bmatrix} v_{1,1} & v_{2,1} & \cdots & v_{D,1} \end{bmatrix} \\ &= \begin{bmatrix} v_{1,1}\mathbf{z}_1 & v_{2,1}\mathbf{z}_1 & \cdots & v_{D,1}\mathbf{z}_1 \end{bmatrix} \end{aligned} \quad (50)$$

观察图 32 热图可以发现一些有意思的特点。还原得到的数据每一列热图模式高度相似；(50) 解释了这一点， \mathbf{X}_1 的每一列均是标量乘以向量 \mathbf{z}_1 的结果。显然， \mathbf{X}_1 的秩为 1，即 $\text{rank}(\mathbf{X}_1) = 1$ 。

图 33、图 34 和图 35 分别展示 \mathbf{z}_2 、 \mathbf{z}_3 和 \mathbf{z}_4 还原 \mathbf{X} 部分数据。

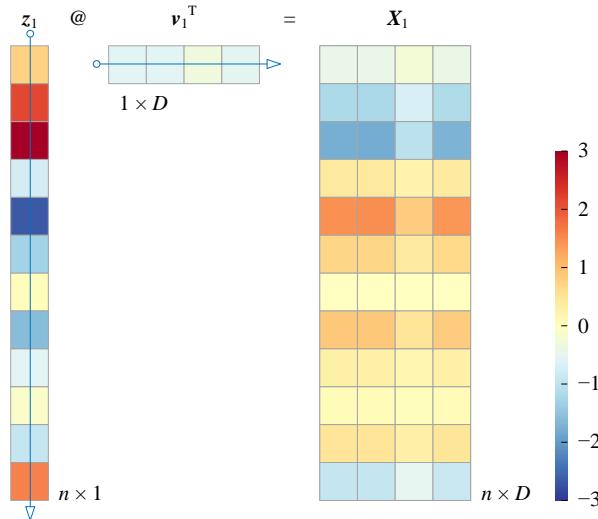


图 32. \mathbf{z}_1 还原 \mathbf{X} 部分数据

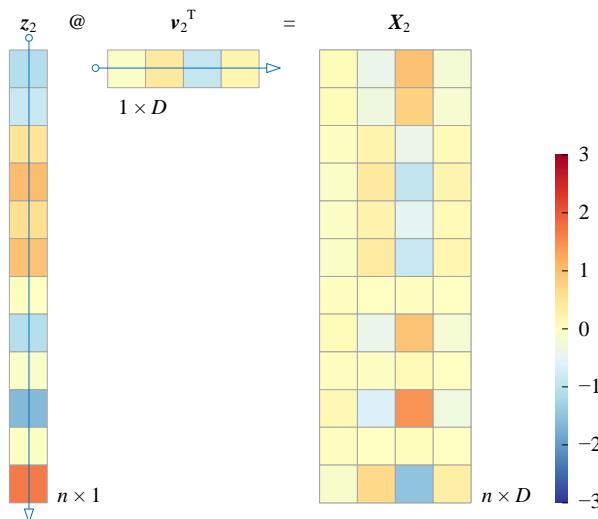


图 33. \mathbf{z}_2 还原 \mathbf{X} 部分数据

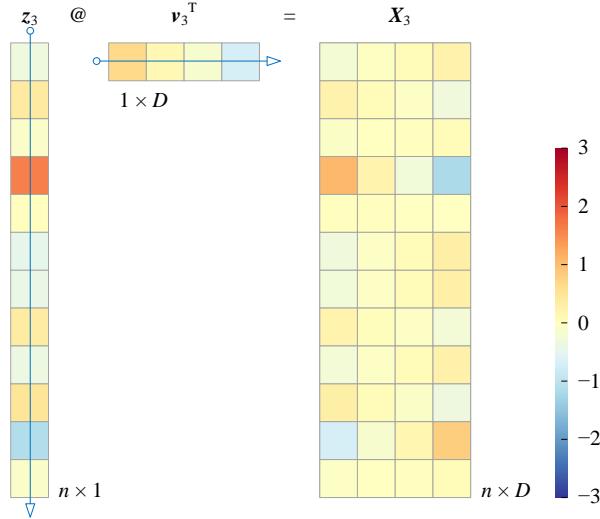
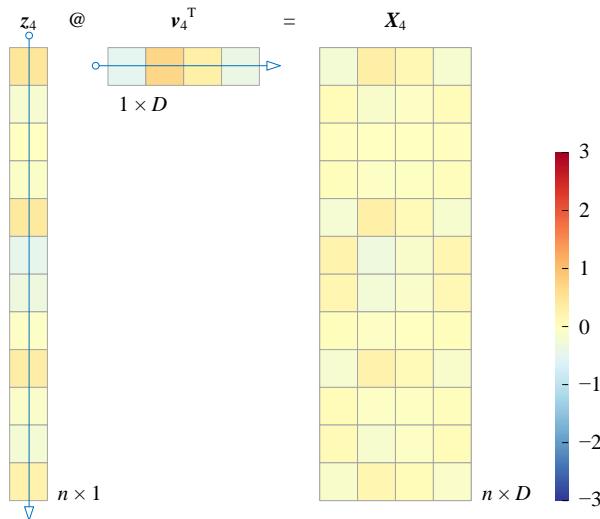
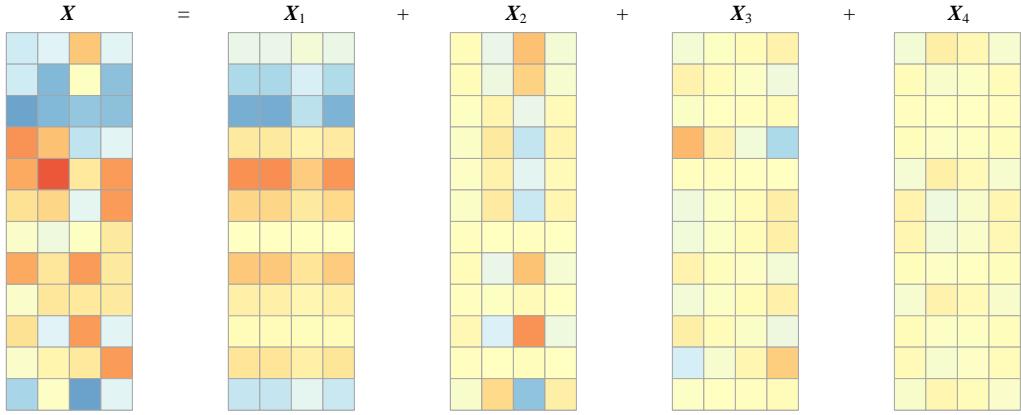
图 34. z_3 还原 X 部分数据图 35. z_4 还原 X 部分数据

图 36 所示为原始数据矩阵 X 热图相当于四层热图叠加结果。观察图 36，发现随着主成分次数降低，每个主成分各自对数据 X 还原力度不断降低，看到还原热图颜色越来越浅；但是，把这些主成分各自还原生成热图不断叠加，获得热图就不断逼近原始热图。

图 36. 原始数据矩阵 \mathbf{X} 热图于四层热图叠加结果

张量积

另外，(48) 可以用张量积来表达：

$$\mathbf{X} = \underbrace{\mathbf{z}_1 \otimes \mathbf{v}_1}_{\hat{\mathbf{X}}_1} + \underbrace{\mathbf{z}_2 \otimes \mathbf{v}_2}_{\hat{\mathbf{X}}_2} + \underbrace{\mathbf{z}_3 \otimes \mathbf{v}_3}_{\hat{\mathbf{X}}_3} + \underbrace{\mathbf{z}_4 \otimes \mathbf{v}_4}_{\hat{\mathbf{X}}_4} \quad (51)$$

利用 (14), (48) 可以整理为：

$$\mathbf{X} = \mathbf{X}\mathbf{v}_1\mathbf{v}_1^T + \mathbf{X}\mathbf{v}_2\mathbf{v}_2^T + \dots + \mathbf{X}\mathbf{v}_D\mathbf{v}_D^T = \sum_{j=1}^D \mathbf{X}\mathbf{v}_j\mathbf{v}_j^T = \mathbf{X} \left(\sum_{j=1}^D \mathbf{v}_j\mathbf{v}_j^T \right) \quad (52)$$

(52) 可以用张量积表达：

$$\mathbf{X} = \mathbf{X}(\mathbf{v}_1 \otimes \mathbf{v}_1) + \mathbf{X}(\mathbf{v}_2 \otimes \mathbf{v}_2) + \dots + \mathbf{X}(\mathbf{v}_D \otimes \mathbf{v}_D) = \sum_{j=1}^D \mathbf{X}\mathbf{v}_j \otimes \mathbf{v}_j = \mathbf{X} \left(\sum_{j=1}^D \mathbf{v}_j \otimes \mathbf{v}_j \right) \quad (53)$$

图 37 所示为通过主成分 $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$ 和其自身转置乘积计算张量积。

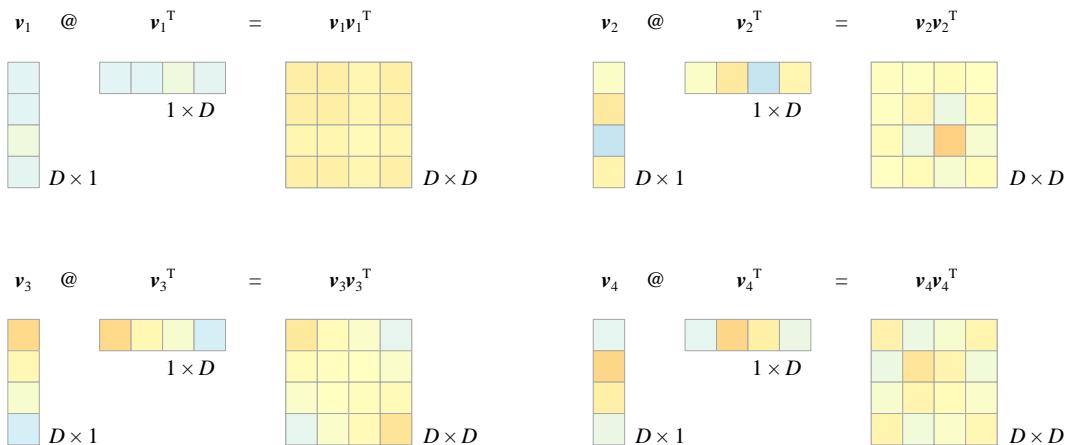


图 37. 列向量乘自身转置获得四个张量积

图 38 所示为张量积运算，和图 37 结果完全一致。

$$\begin{array}{ccc}
 \mathbf{v}_1 & \otimes & \mathbf{v}_1 = \mathbf{v}_1 \otimes \mathbf{v}_1 \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 D \times 1 & D \times 1 & D \times D \\
 \end{array} \quad
 \begin{array}{ccc}
 \mathbf{v}_2 & \otimes & \mathbf{v}_2 = \mathbf{v}_2 \otimes \mathbf{v}_2 \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 D \times 1 & D \times 1 & D \times D \\
 \end{array} \\
 \\
 \begin{array}{ccc}
 \mathbf{v}_3 & \otimes & \mathbf{v}_3 = \mathbf{v}_3 \otimes \mathbf{v}_3 \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 D \times 1 & D \times 1 & D \times D \\
 \end{array} \quad
 \begin{array}{ccc}
 \mathbf{v}_4 & \otimes & \mathbf{v}_4 = \mathbf{v}_4 \otimes \mathbf{v}_4 \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \\
 D \times 1 & D \times 1 & D \times D \\
 \end{array}$$

图 38. 内积计算获得四个张量积

容易推导得到，(53) 中张量积相加得到单位矩阵：

$$\mathbf{v}_1 \otimes \mathbf{v}_1 + \mathbf{v}_2 \otimes \mathbf{v}_2 + \dots + \mathbf{v}_D \otimes \mathbf{v}_D = \left(\sum_{j=1}^D \mathbf{v}_j \otimes \mathbf{v}_j \right) = \mathbf{I} \quad (54)$$

上式如图 39 热图所示。

$$\begin{array}{ccccccccc}
 \mathbf{v}_1 \otimes \mathbf{v}_1 & + & \mathbf{v}_2 \otimes \mathbf{v}_2 & + & \mathbf{v}_3 \otimes \mathbf{v}_3 & + & \mathbf{v}_4 \otimes \mathbf{v}_4 & = & \mathbf{I} \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \\
 \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \begin{array}{c} \text{---} \\ \text{---} \\ \text{---} \\ \text{---} \end{array} & & \\
 D \times D & & D \times D \\
 \end{array}$$

图 39. 张量积相加得到单位矩阵

联立 (15) 和 (49)，利用张量积 $\mathbf{v}_1 \otimes \mathbf{v}_1$ 还原部分原始数据：

$$\mathbf{X}_1 = \mathbf{z}_1 \mathbf{v}_1^\top = \mathbf{X} \mathbf{v}_1 \mathbf{v}_1^\top = \mathbf{X} \underbrace{(\mathbf{v}_1 \otimes \mathbf{v}_1)}_{\text{Tensor product}} \quad (55)$$

类似，张量积 $\mathbf{v}_2 \otimes \mathbf{v}_2$ 也可以还原部分原始数据：

$$\mathbf{X}_2 = \mathbf{z}_2 \mathbf{v}_2^\top = \mathbf{X} \mathbf{v}_2 \mathbf{v}_2^\top = \mathbf{X} \underbrace{(\mathbf{v}_2 \otimes \mathbf{v}_2)}_{\text{Tensor product}} \quad (56)$$

图 40 所示为张量积 $\mathbf{v}_1 \otimes \mathbf{v}_1$ 和 $\mathbf{v}_2 \otimes \mathbf{v}_2$ 还原部分数据 \mathbf{X} ；图 41 所示为张量积 $\mathbf{v}_3 \otimes \mathbf{v}_3$ 和 $\mathbf{v}_4 \otimes \mathbf{v}_4$ 还原部分数据 \mathbf{X} 。



《矩阵力量》第 10 章给这种投影一个特别的名字——二次投影，建议大家回顾。

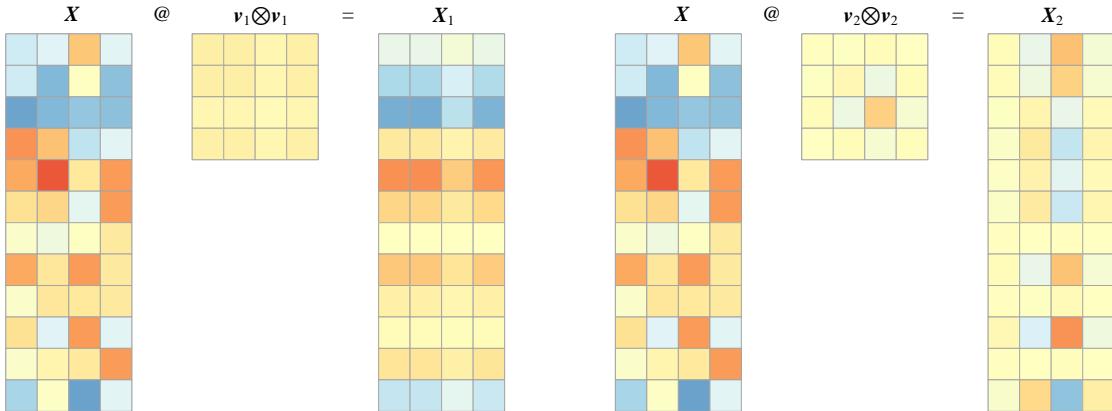


图 40. 张量积 $X(v_1 \otimes v_1)$ 和 $X(v_2 \otimes v_2)$ 还原部分数据 X

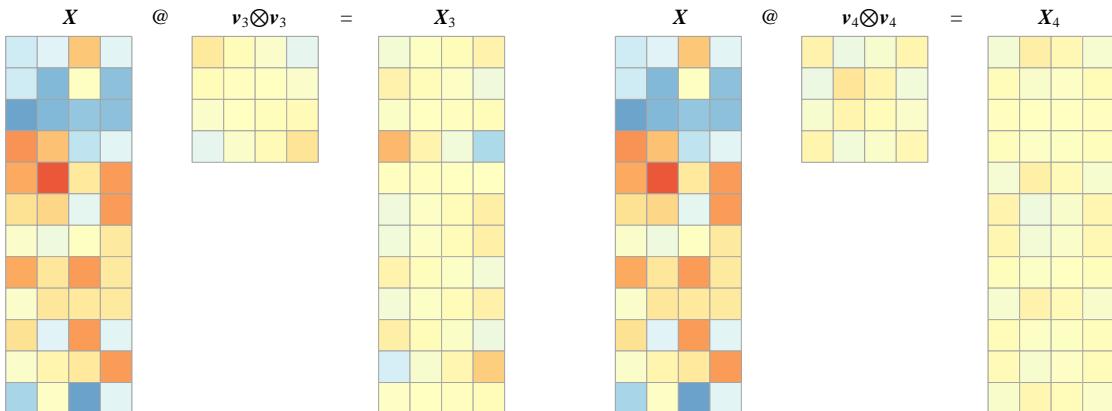


图 41. 张量积 $X(v_3 \otimes v_3)$ 和 $X(v_4 \otimes v_4)$ 还原部分数据 X

误差

图 42 所示为两个主成分 v_1 和 v_2 还原获得原始数据热图，具体计算如下：

$$\hat{X} = [z_1 \ z_2] [v_1 \ v_2]^T \quad (57)$$

相当于

$$\begin{aligned} \hat{X} &= X_1 + X_2 = z_1 v_1^T + z_2 v_2^T \\ &= X (v_1 v_1^T + v_2 v_2^T) = X (v_1 \otimes v_1 + v_2 \otimes v_2) \end{aligned} \quad (58)$$

图 43 所示为通过叠加图 32 和图 33 两个热图还原原始数据矩阵。

从张量积角度来看图 43，

$$\mathbf{X} \approx \mathbf{X} (\mathbf{v}_1 \otimes \mathbf{v}_1 + \mathbf{v}_2 \otimes \mathbf{v}_2) = s_1 \mathbf{u}_1 \otimes \mathbf{v}_1 + s_2 \mathbf{u}_2 \otimes \mathbf{v}_2^T \quad (59)$$

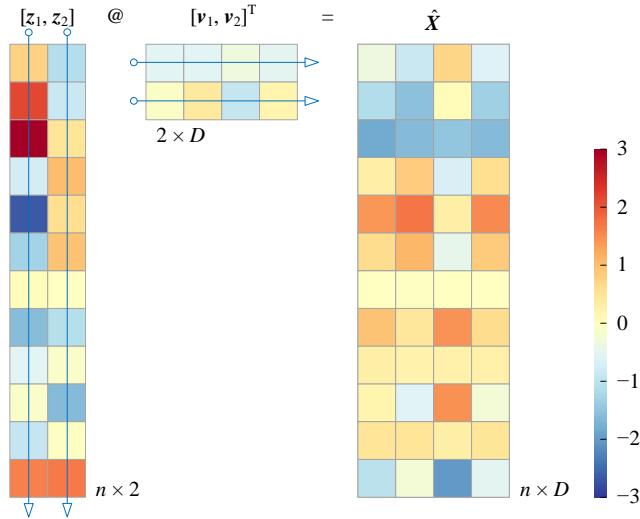
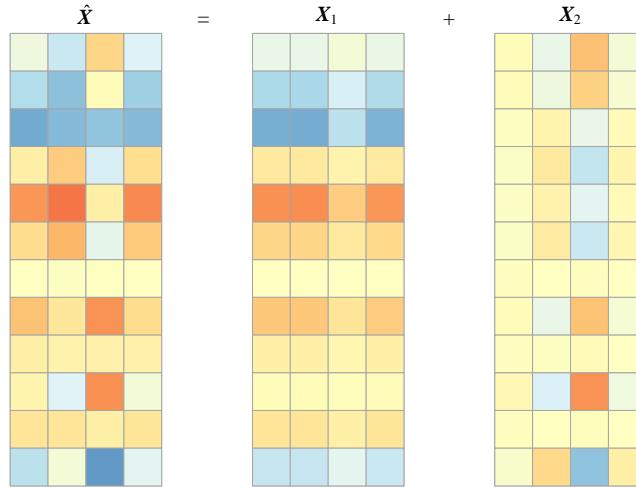
图 42. 前两个主成分 z_1 和 z_2 还原 X 数据

图 43. 两个热图叠加还原原始数据

残差数据矩阵 \mathbf{E} , 即原始热图和还原热图色差, 利用下式计算获得:

$$\mathbf{E} = \mathbf{X} - \hat{\mathbf{X}} \quad (60)$$

图 44 比较原始数据 \mathbf{X} 、拟合数据 $\hat{\mathbf{X}}$ 和残差数据矩阵 \mathbf{E} 热图, 发现原始数据 \mathbf{X} 和拟合数据 $\hat{\mathbf{X}}$ 已经相差无几。从图片还原角度来看, 如图 44 所示, PCA 降维用更少维度、更少数据获得几乎一样画质图片。

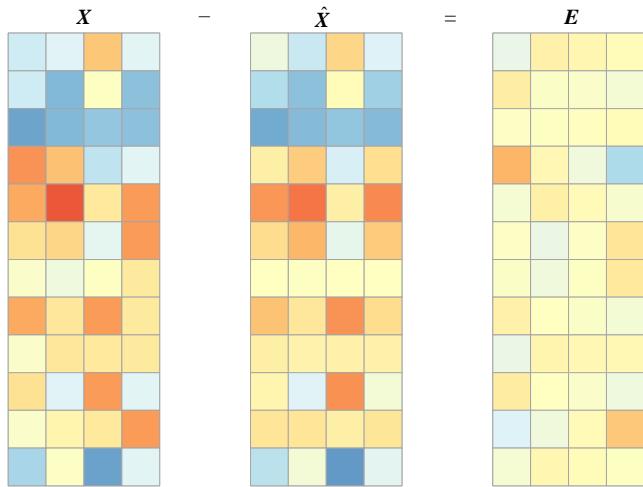


图 44. 原始数据、拟合数据和残差数据热图

六条技术路径

相信大家对表 1 并不陌生，大家都在《矩阵力量》第 25 章中见过这六条 PCA 技术路线。本章介绍的实际上是：a) 特征值分解协方差矩阵；b) 奇异值分解中心化数据矩阵。

总结来说，通过 PCA 降维，我们可以减少数据的维度，从而简化模型和算法的复杂度，同时可以去除噪声和冗余信息，提高数据的可解释性和可视化效果，从而更好地理解数据和发现数据中的规律。PCA 广泛应用于数据挖掘、模式识别、图像处理、信号处理等领域。



《数据有道》一册将比较表 1 这六种方法的异同。

表 1. 六条 PCA 技术路线，来自《矩阵分解》第 25 章

对象	方法	结果
原始数据矩阵 X	奇异值分解	$X = U_X S_X V_X^T$
格拉姆矩阵 $G = X^T X$ 本章中用“修正”的格拉姆矩阵 $G = \frac{X^T X}{n-1}$	特征值分解	$G = V_X A_X V_X^T$
中心化数据矩阵 $X_c = X - E(X)$	奇异值分解	$X_c = U_c S_c V_c^T$
协方差矩阵 $\Sigma = \frac{(X - E(X))^T (X - E(X))}{n-1}$	特征值分解	$\Sigma = V_c A_c V_c^T$
标准化数据 (z 分数) $Z_x = (X - E(X)) D^{-1}$ $D = \text{diag}(\text{diag}(\Sigma)^{\frac{1}{2}})$	奇异值分解	$Z_x = U_z S_z V_z^T$

$P = D^{-1} \Sigma D^{-1}$ 相关性系数矩阵 $D = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}}$	特征值分解	$P = V \Lambda z V^T$
---	-------	-----------------------



人类思维天然具备概率统计属性。概率统计的背后的思想更贴近“生活常识”。大脑涉及可能性判断时，就不自觉进入“贝叶斯推断”模式。

看着天上云层很厚，可能两小时就会下雨。昨晚淋了雨，估计今天要感冒。根据以往经验，估计这次考试通过率 80% 以上。这种“先验 + 数据 → 后验”的思维模式比比皆是。

可惜的是，当数学家将这些生活常识“翻译成”数学语言之后，它们就变成了冷冰冰“火星文”。

概率统计与其说是工具，不如说是方法论、世界观。大家常说的“一命，二运，三风水，四读书”，体现的也是概率统计的思维。

天意从来高难问，命中没有莫强求。“小概率事件”能发生，得之我幸，不得我命。风水轮流转，玄而又玄。

目不转睛地盯着社会财富分布曲线的“右尾”，对巨贾兜售的“成功学”布道言听计从，从统计角度来看都是痴人说梦。科技巨头退学创业的成功“典范”对应的概率也不比“买彩票中头奖”高多少。

正所谓知识改变命运，只有读书成才对应“大概率事件”。大家捧起“鸢尾花书”的时候，就依靠统计思维做出了“优化”选择。

《统计至简》是“鸢尾花书”数学板块的三本中的最后一本，其中大家看到了代数、几何、线性代数、概率统计、优化等数学板块的合流。

读到这，大家便完成了整个数学板块的修炼。希望大家日后再看到任何公式的时候，闭上眼睛，都能在脑中“看见”各种几何图形。

还有，让我们和鸡、兔、猪这三个伙伴说声感谢！感谢它们在学习数学路上的陪伴！再见，为了下次更好的遇见！

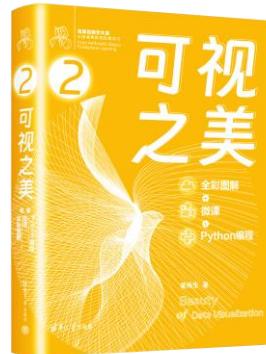
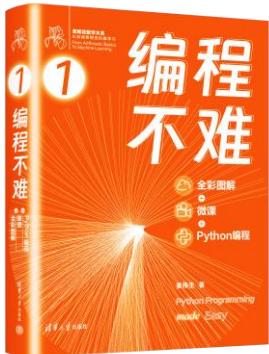
下面，我们一起踏上《数据有道》、《机器学习》的“实践”之旅！



“鸢尾花书”的整体布局

数学

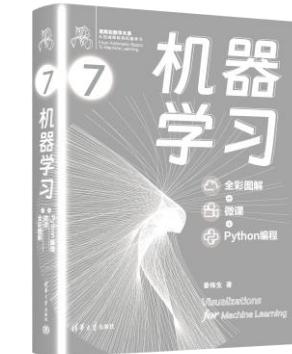
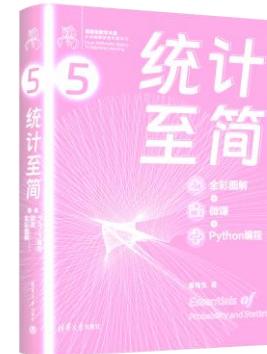
数学基础



线性代数



概率统计



Python编程

数据可视化

回归、降维

分类、聚类

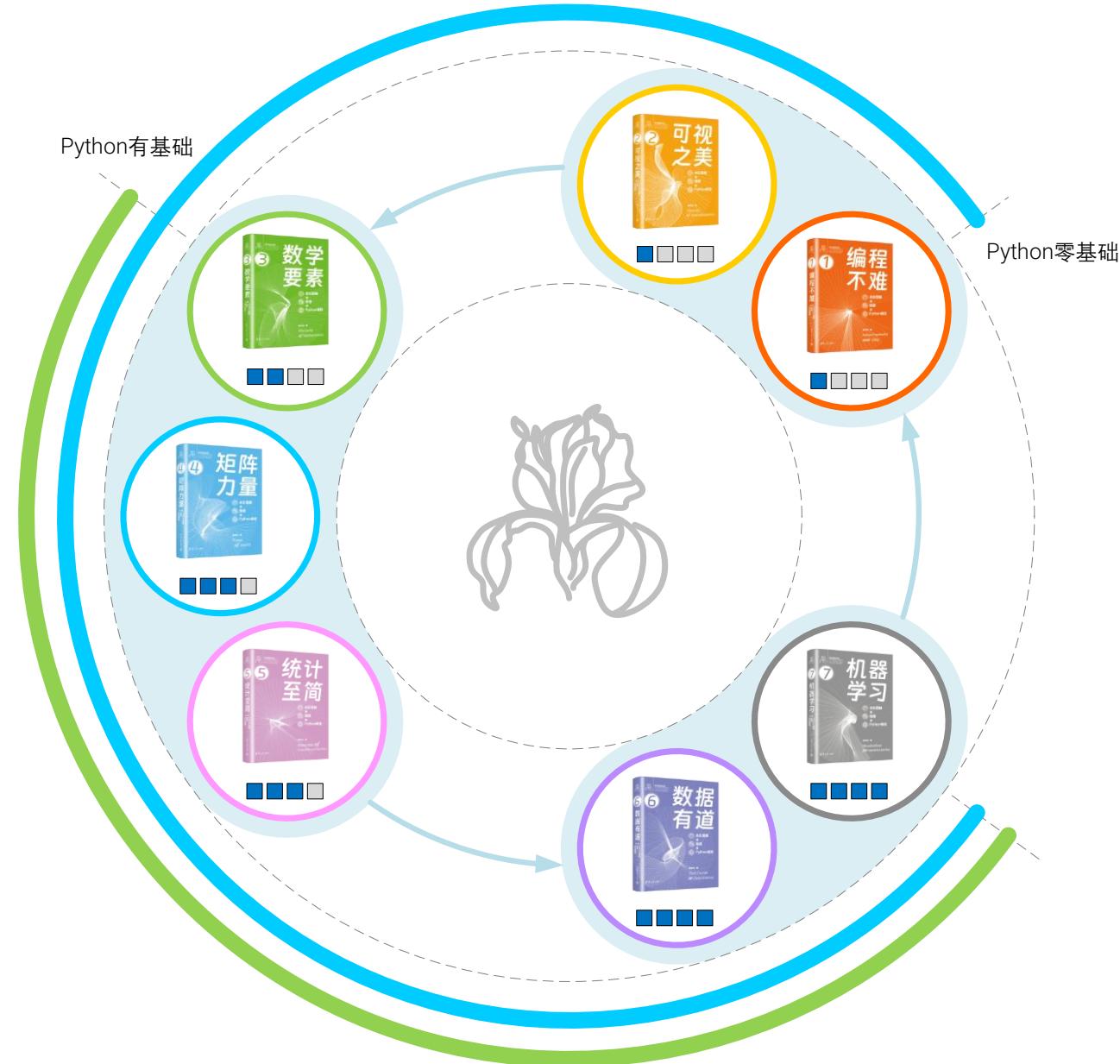
工具

实践

数学 + Python编程 + 可视化 + 机器学习实践



“鸢尾花书”的学习顺序

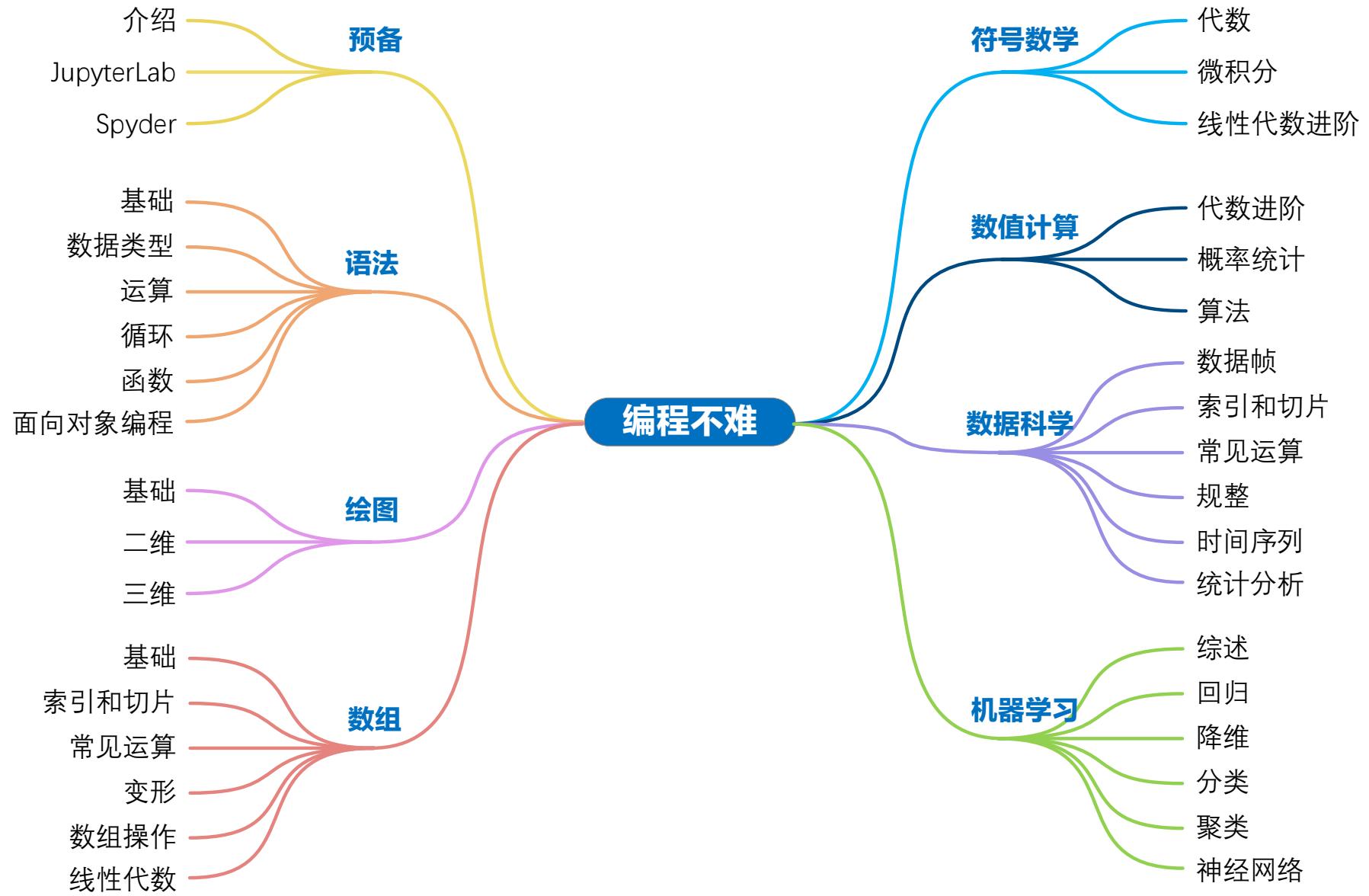


分册进度状态

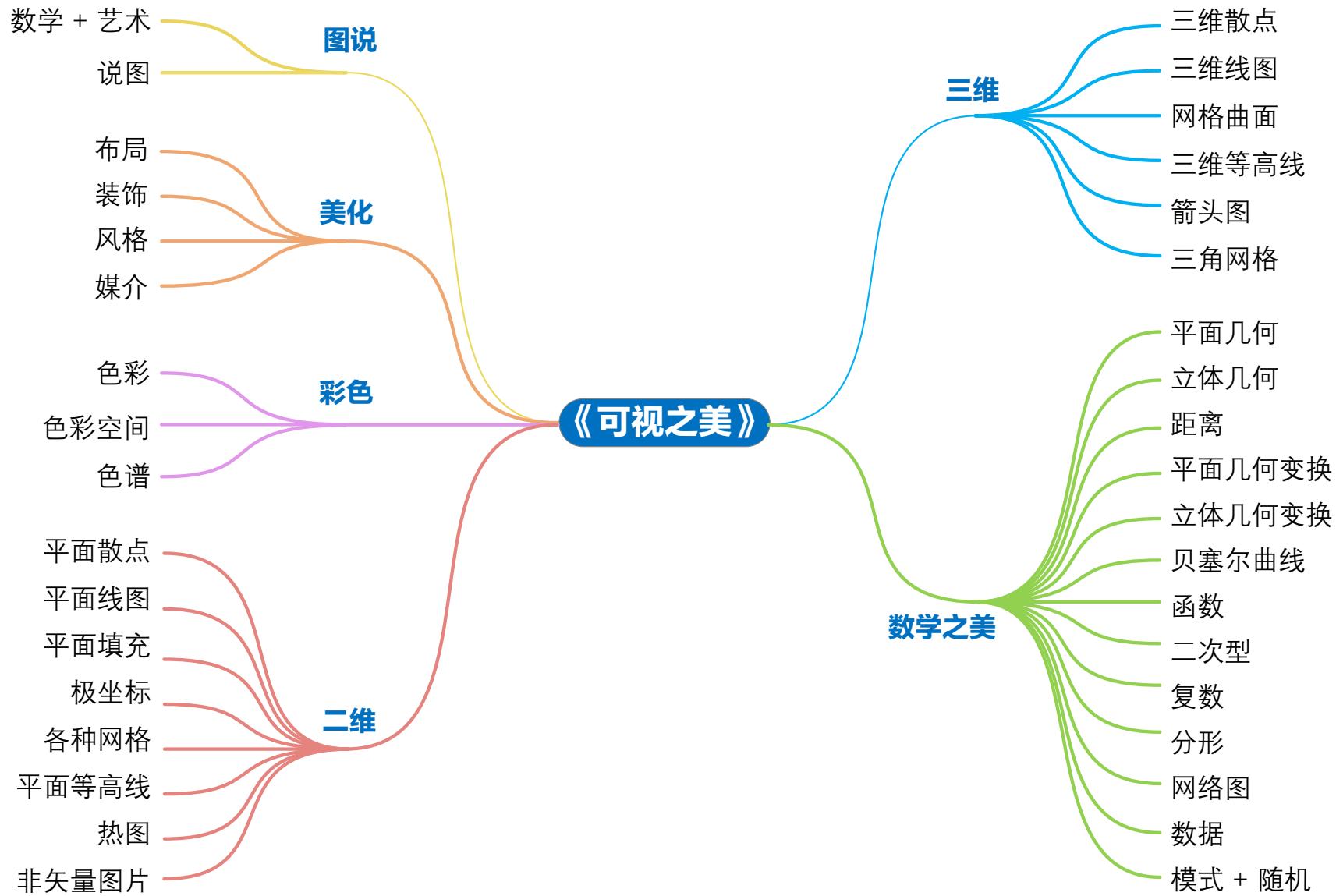
	草稿、Python	打磨、视频	清华社五审五校	上架
1 编程不难		~50%		
2 可视之美		~100%	2023, 10	
3 数学要素		100%	完成	完成
4 矩阵力量		100%	完成	完成
5 统计至简		100%	2023, 07	2023, 08
6 数据有道		80%	2024年初	
7 机器学习		80%	2024年初	



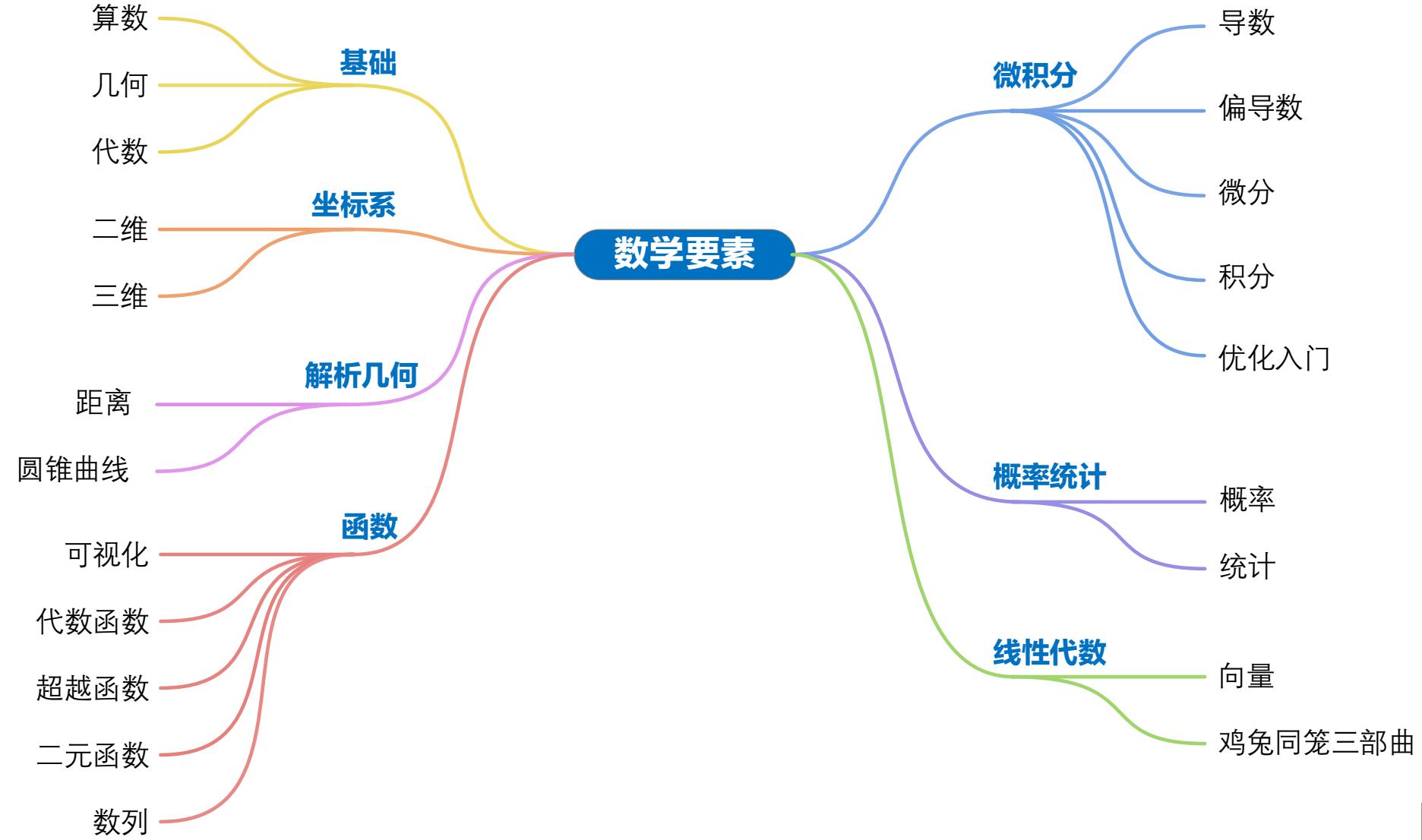
Book 1 《编程不难》



Book 2 《可视之美》



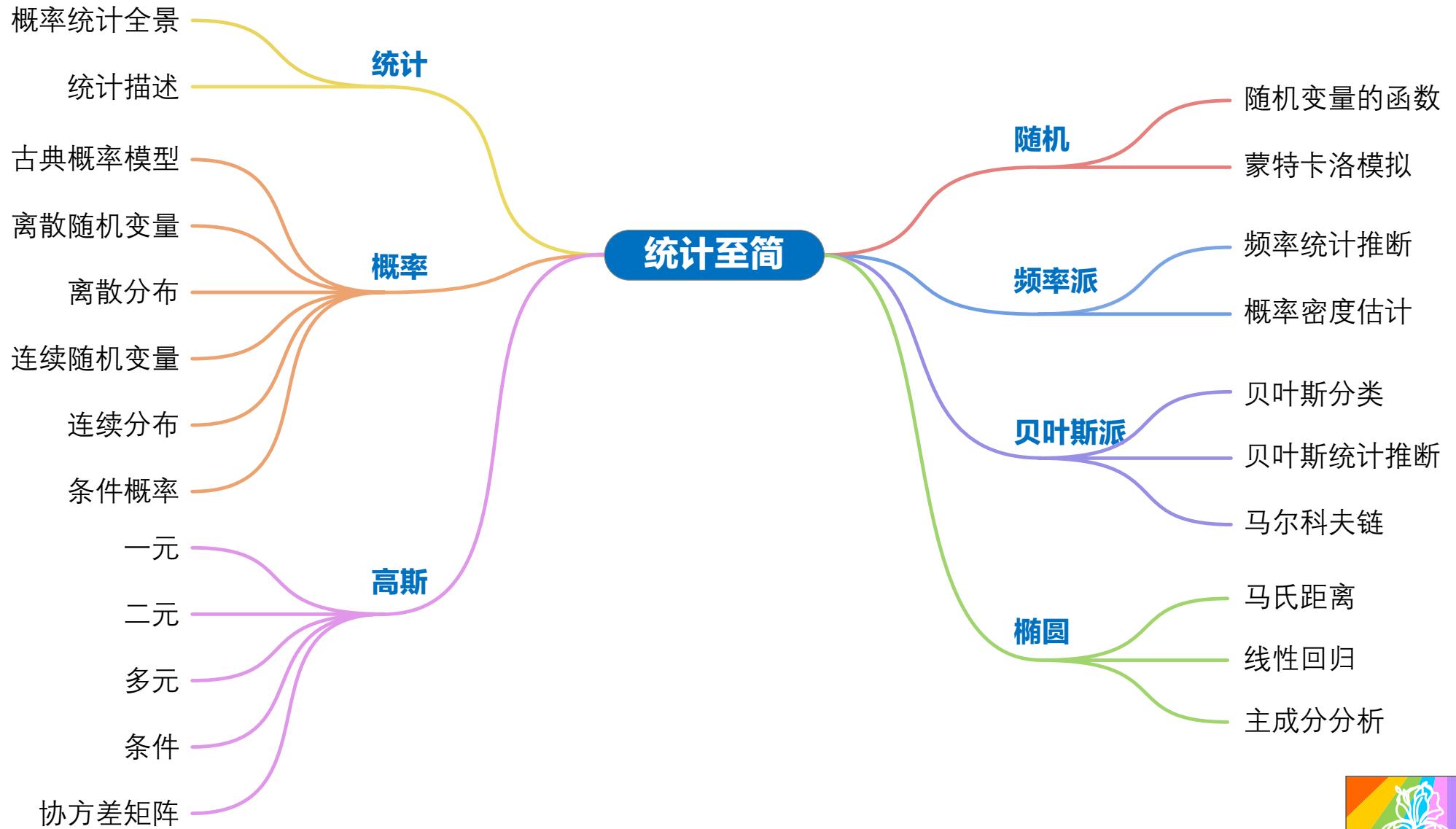
Book 3 《数学要素》



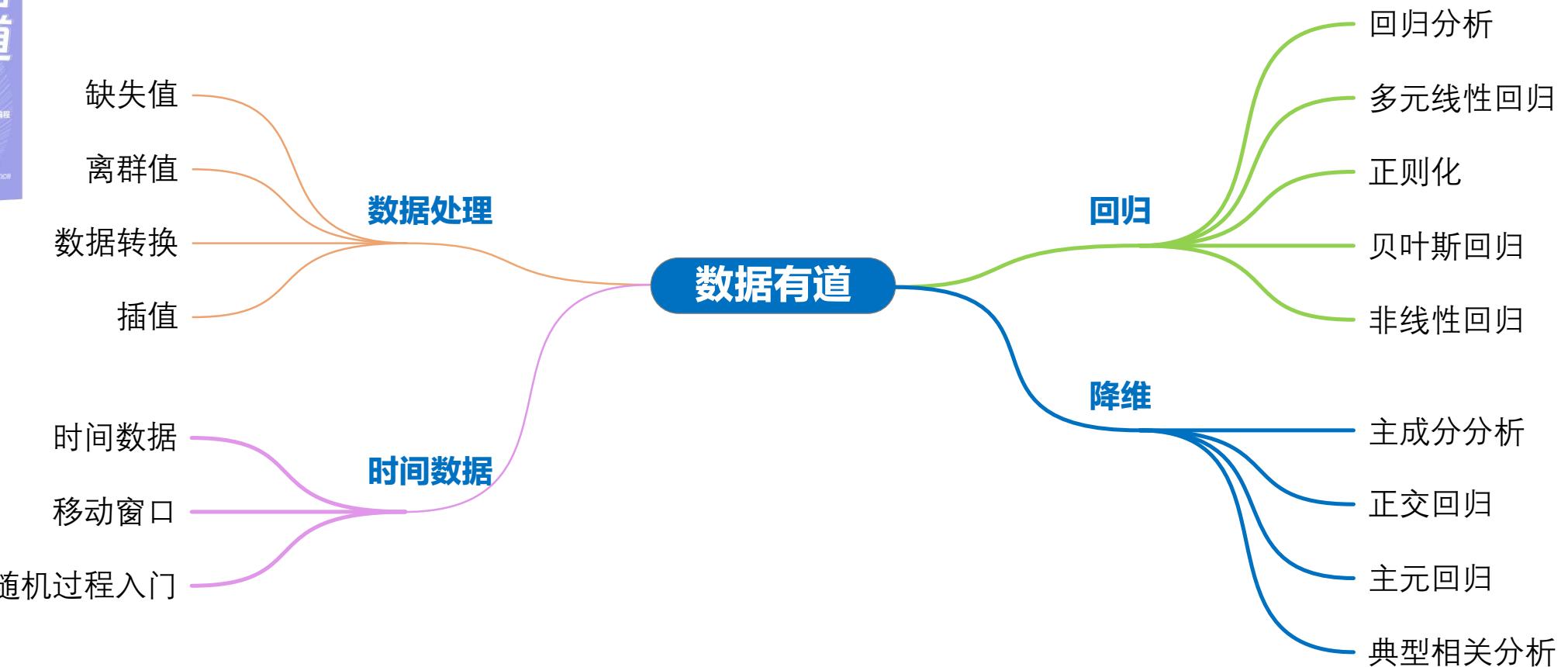
Book 4 《矩阵力量》



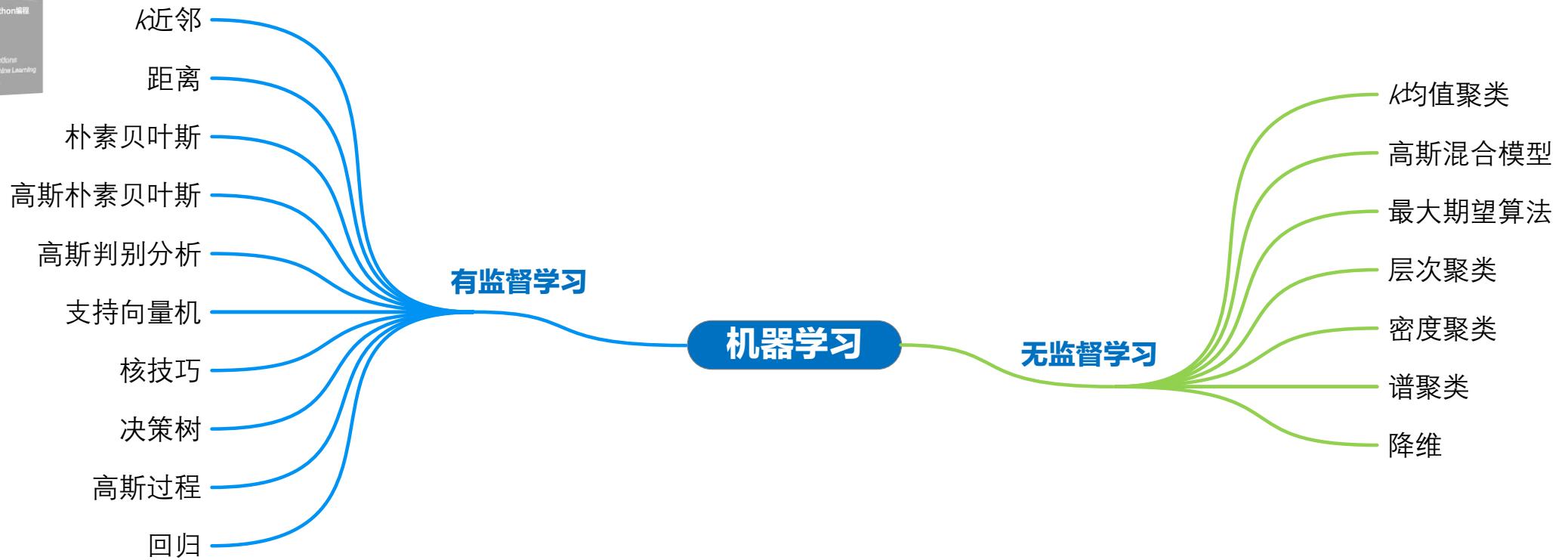
Book 5 《统计至简》



Book 6 《数据有道》



Book 7 《机器学习》



开源资源

PDF书稿、代码: <https://github.com/Visualize-ML>

微课视频: <https://space.bilibili.com/513194466>

信息发布: <https://www.zhihu.com/people/jamestong-xue>

专属邮箱: jiang.visualize.ml@gmail.com

