

2

Let's Talk about Figures

说图

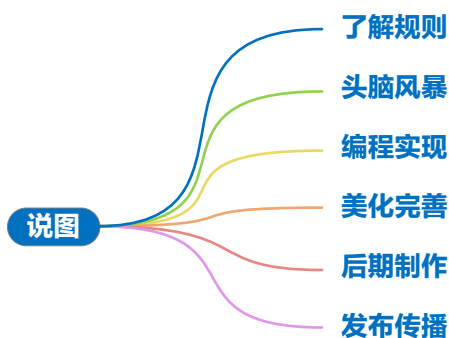
正式开始一场“数学 + 艺术”的动手实践



独处，是创造的秘诀；独处，是好主意诞生的时候。

Be alone, that is the secret of invention; be alone, that is when ideas are born.

—— 尼古拉·特斯拉 (Nikola Tesla) | 发明家、物理学家 | 1856 ~ 1943



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 一图胜千言

上一章，我们聊了一些有关“数学 + 艺术”形而上务虚的内容，本章开始介绍如何用 Python 完成各种可视化的实操内容。

一图胜千言 (a picture is worth a thousand words)。一说到图片可视化的作用，大家自然而然地会想到这句“陈词滥调”。但是，并不是所有的图片都“胜千言”。

可能在某些场合中“颜值即正义”，但是在数学工具、数据科学、机器学习等应用场景，优质可视化方案不仅仅要读者“眼前一亮”且“言之有物”。

优质可视化方案有助于高效传播信息，在短时间内让读者接收到信息，并促进交流、思考。有效的图片信息传播应注重高颜值、清晰度、专业性、简洁性、准确性和与读者之间的互动。简而言之，运用之妙，存乎一心，让读者主动思考的图片才是优质的可视化方案。

反之，低效可视化方案问题可能涉及：信息密度低、信息过载（比如满纸公式）、图像质量差（非矢量、分辨率低）、设计混乱（中心不明确、分散注意力）、缺乏明确的标签和解释、配色方案失效、缺乏上下文和交互性，以及信息的不准确性（比如手绘高斯函数曲线或曲面）。

小姜，作者说的信息密度低，是几个意思？



就是咱俩啊！
拉咱俩出来做反例，五官都懒得画 ...

图 1. 低效信息传播

以图 2 为例，为了版面设计，在文本中插入这张鸢尾花照片，图片的作用仅仅是凑数的“花瓶”。如果利用这幅鸢尾花照片讲解一幅彩色照片可以由红、绿、蓝三色组成，这张照片就成了故事链重要的一部分。为了保持信息传播的连续、高效，我们还可以用这幅鸢尾花照片讲解矩阵（图 3）、主成分分析（图 4）等等。



图 2. 鸢尾花照片

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

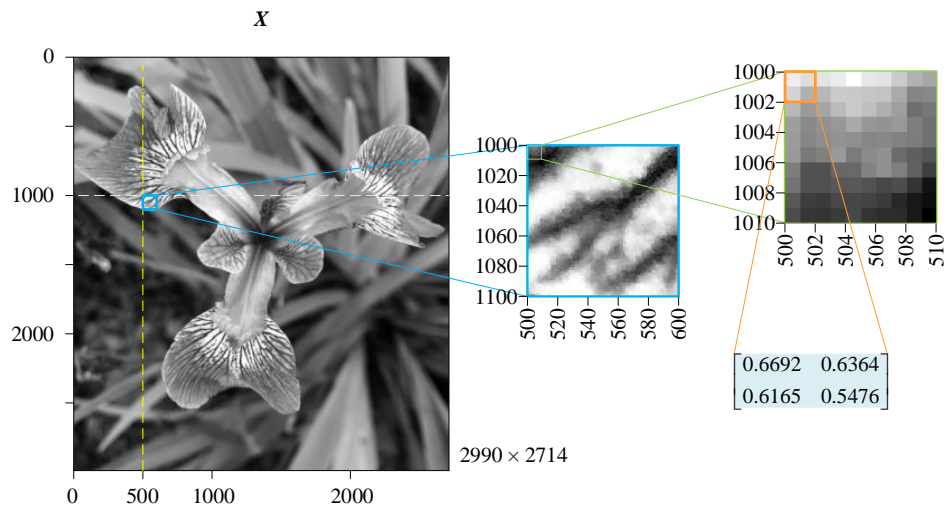


图 3. 照片也是数据矩阵

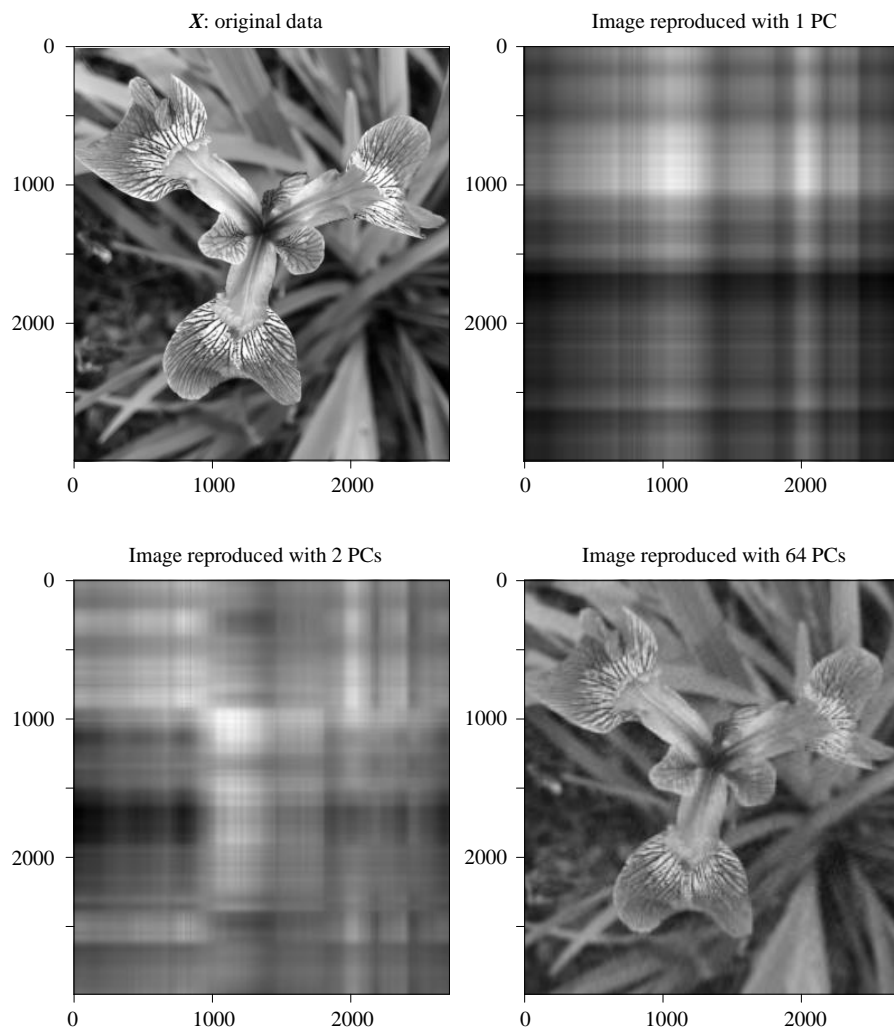


图 4. 对黑白鸢尾花照片的主成分分析

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

一张图片的整个生命周期一般要经过如下几个阶段：

- ▶ 了解规则；
- ▶ 头脑风暴；
- ▶ 编程实现；
- ▶ 美化完善（代码层面）；
- ▶ 后期制作；
- ▶ 发布传播。

下面，我们便按这个顺序介绍如何制作一张图片。

2.2 了解规则：带着枷锁舞蹈

在开始可视化之前，务必明确图片的目标是什么，以及确定图片的受众是谁。不同的目标和受众可能需要不同类型和风格的可视化呈现。

在这个纸媒和数字媒体共存、共荣的多媒体时代，制作一张图片通常要同时照顾到纸媒、数字媒体的需求。

本书介绍的可视化是在科技制图的范畴之内。因此创作一张图片之前，首先关注制图规则，以便决定图片各种属性。

建议大家在创作图片时考虑以下几个问题：

- ▶ 风格是学术专业？还是轻松活泼？
- ▶ 是否允许手绘？
- ▶ 图片大小尺寸？比例如何？
- ▶ 一幅图是否可以多子图？子图布局有何要求？
- ▶ 图片内文字字体（Times New Roman, Arial, Roboto, ...）？
- ▶ 文字字号最大、最小几号？文字颜色是否有要求？
- ▶ 图片中的文字是否要求可编辑？
- ▶ 图片中是否可以嵌入公式？
- ▶ 黑白、彩色？配色有何特殊要求？
- ▶ 是否需要考虑彩色图片在黑白灰打印时呈现效果？
- ▶ 是否需要针对色盲群体调整配色？
- ▶ 颜色采用 RGB，还是 CMYK？
- ▶ 图中线宽、线型是否有要求？
- ▶ 是否有必要删除隐藏图层的元素？

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

- ▶ 图片是静态，还是交互？
- ▶ 图片的格式？矢量图，还是像素图？
- ▶ 图片如果过大，是否可以光栅化（**rasterize**）？
- ▶ 像素图的像素要求如何？最小、最大像素？
- ▶ 图片是否需要单独保存，并提交？
- ▶ 图片文件格式（JPEG、PNG、GIF、SVG、TIFF、PDF ...）、大小是否有要求？
- ▶ 图片是否要用于演示，比如放在 PPT 中？PPT 中的文字大小如何？插图文字大小如何？
- ▶ 是否需要制作动画，比如 GIF？
- ▶ 是否考虑创作 App 应用、dashboard？

正式出版物（纸媒、数字媒体、会议）一般都有专门的制图指南，建议大家在开始创作任何图片之前首先仔细阅读制图指南的细则。

如果找不到相关的制图指南，建议大家参考《自然》杂志的制图指南，链接如下：

https://www.nature.com/documents/Final_guide_to_authors.pdf

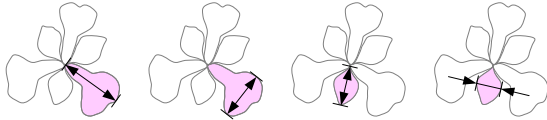
鸢尾花书在创作时，很多制图细节都参考了《自然》的制图指南。

2.3 头脑风暴：知识网络

富有创意的可视化方案可以为数据插上“翅膀”！根据你要传达的信息和数据的性质，选择适合的图表类型，从而提高可视化的效果和可读性。

图 5 所示为鸢尾花数据集，图 17 所示为以鸢尾花数据为起点的知识网。图 6 ~ 图 15 为从鸢尾花书各册精选出来和鸢尾花数据集有关的可视化方案。本书不会介绍这些图背后的数学原理，但是会和大家探讨如何完成这些可视化方案。

图 6 ~ 图 15 都离不开 Python 编程实现，下面简单介绍 Python 在可视化中的作用。



Index	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Species C
1	5.1	3.5	1.4	0.2	Setosa C_1
2	4.9	3	1.4	0.2	
3	4.7	3.2	1.3	0.2	
...	
49	5.3	3.7	1.5	0.2	
50	5	3.3	1.4	0.2	Versicolor C_2
51	7	3.2	4.7	1.4	
52	6.4	3.2	4.5	1.5	
53	6.9	3.1	4.9	1.5	
...	
99	5.1	2.5	3	1.1	Virginica C_3
100	5.7	2.8	4.1	1.3	
101	6.3	3.3	6	2.5	
102	5.8	2.7	5.1	1.9	
103	7.1	3	5.9	2.1	
...	
149	6.2	3.4	5.4	2.3	
150	5.9	3	5.1	1.8	




图 5. 鸢尾花数据表格，单位为厘米 (cm)

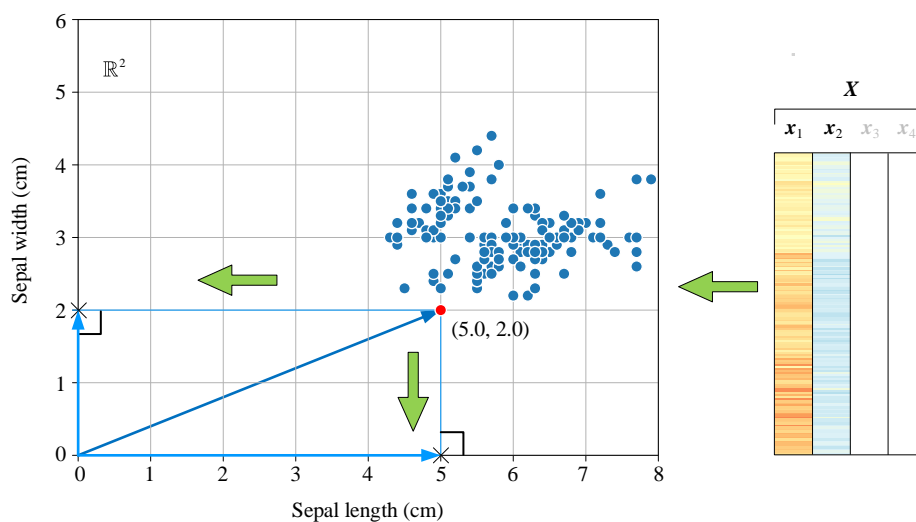


图 6. 鸢尾花前两个特征数据散点图

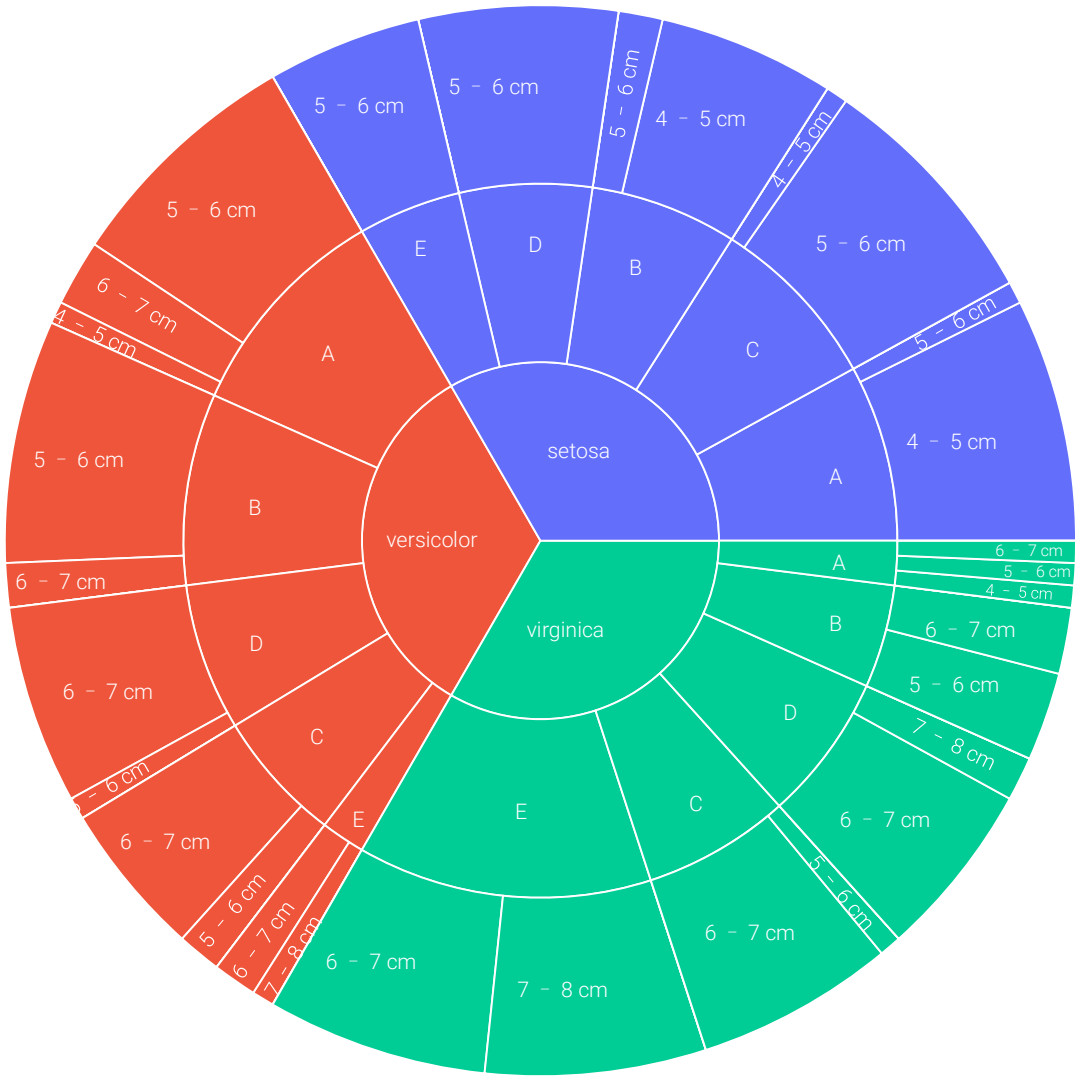


图 7. 太阳爆炸图完成鸢尾花数据钻取

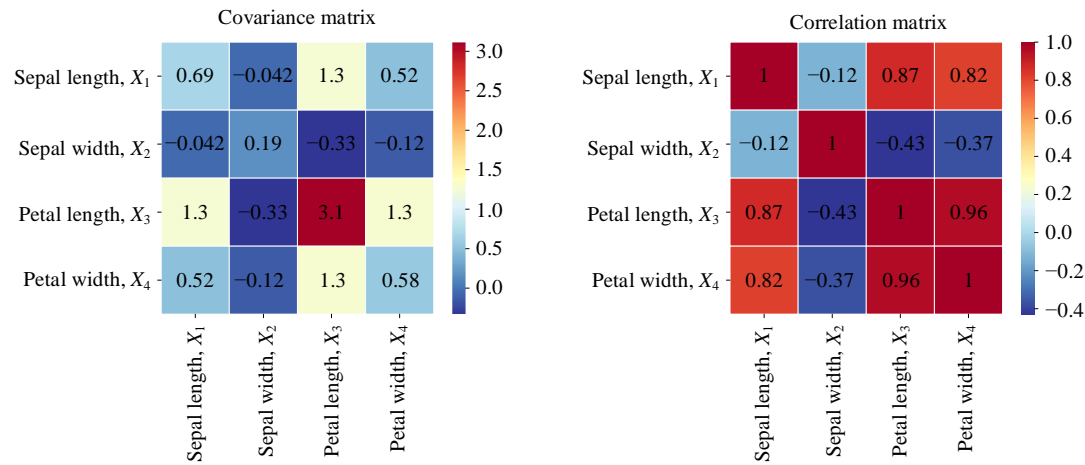


图 14. 协方差矩阵和相关性系数矩阵热图

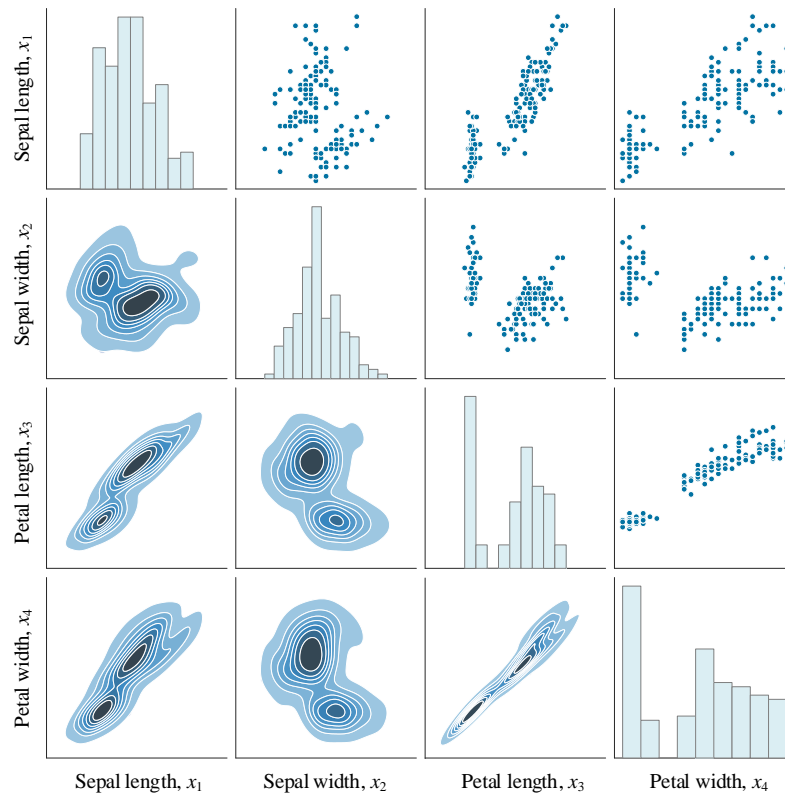


图 8. 鸢尾花数据成对特征分析图，不分类

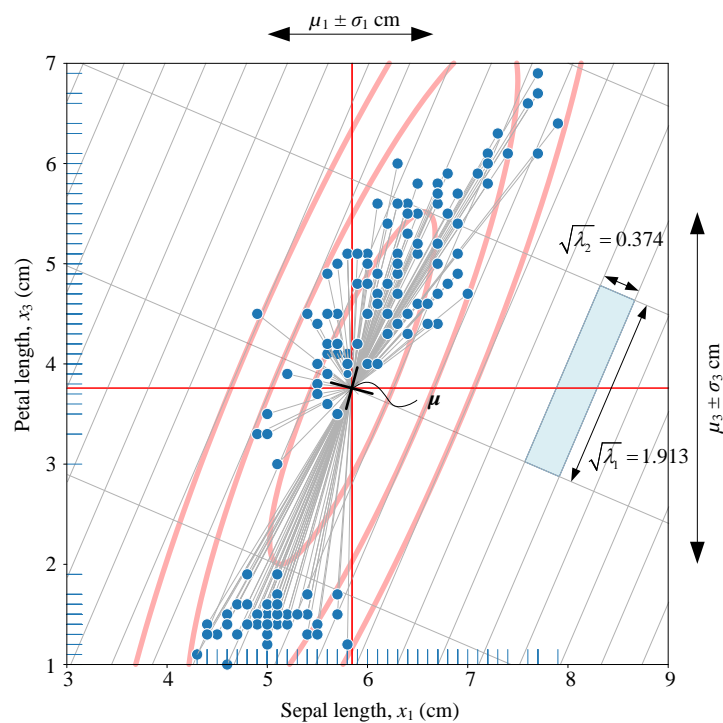


图 9. 花萼长度、花瓣长度平面上的马氏距离等高线和网格

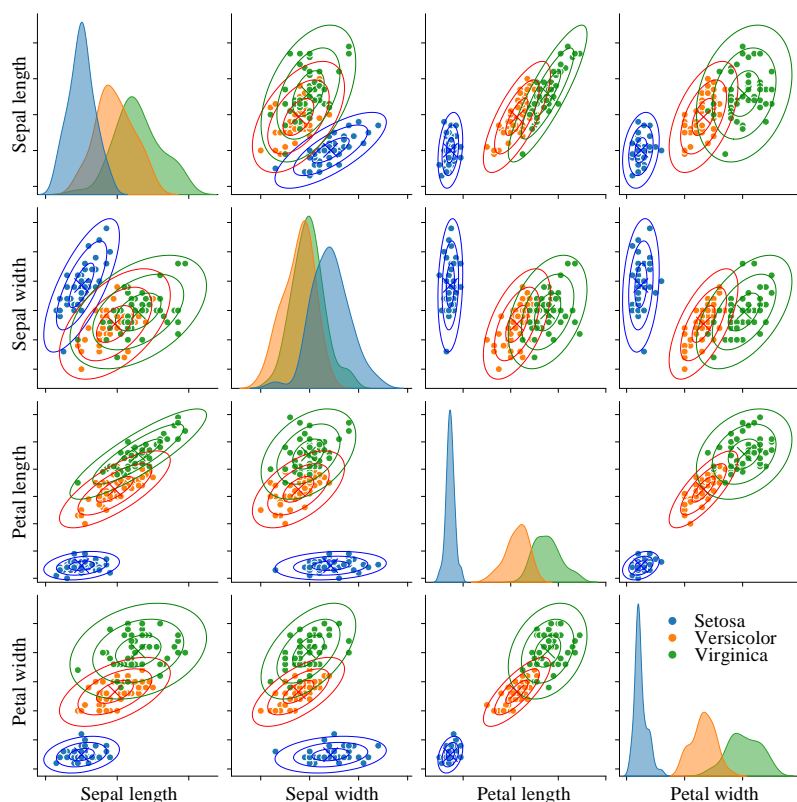
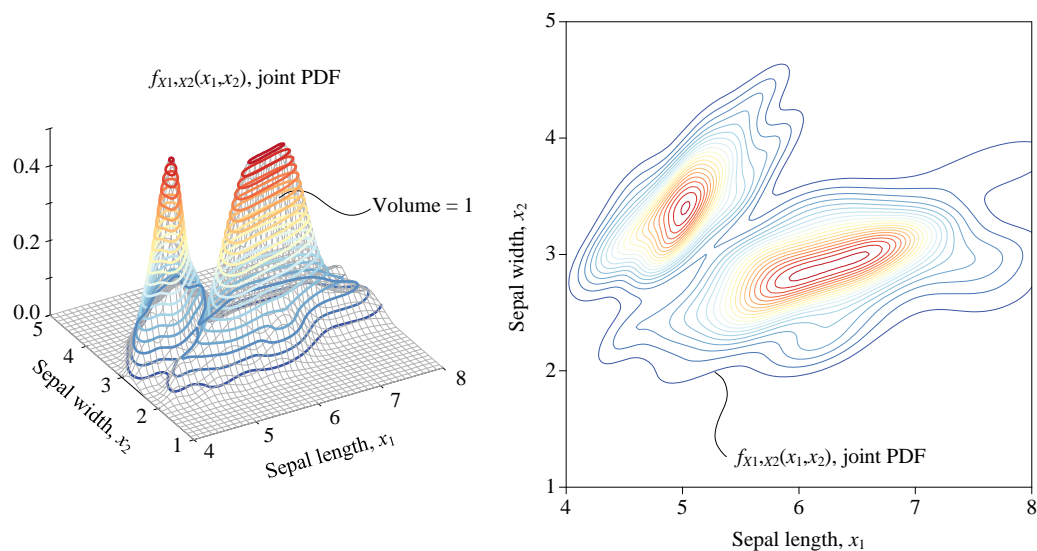


图 18. 协方差矩阵和椭圆的关系，考虑分类

图 10. 联合概率密度函数 $f_{X1,X2}(x_1, x_2)$ 三维等高线和平面等高线，不考虑分类

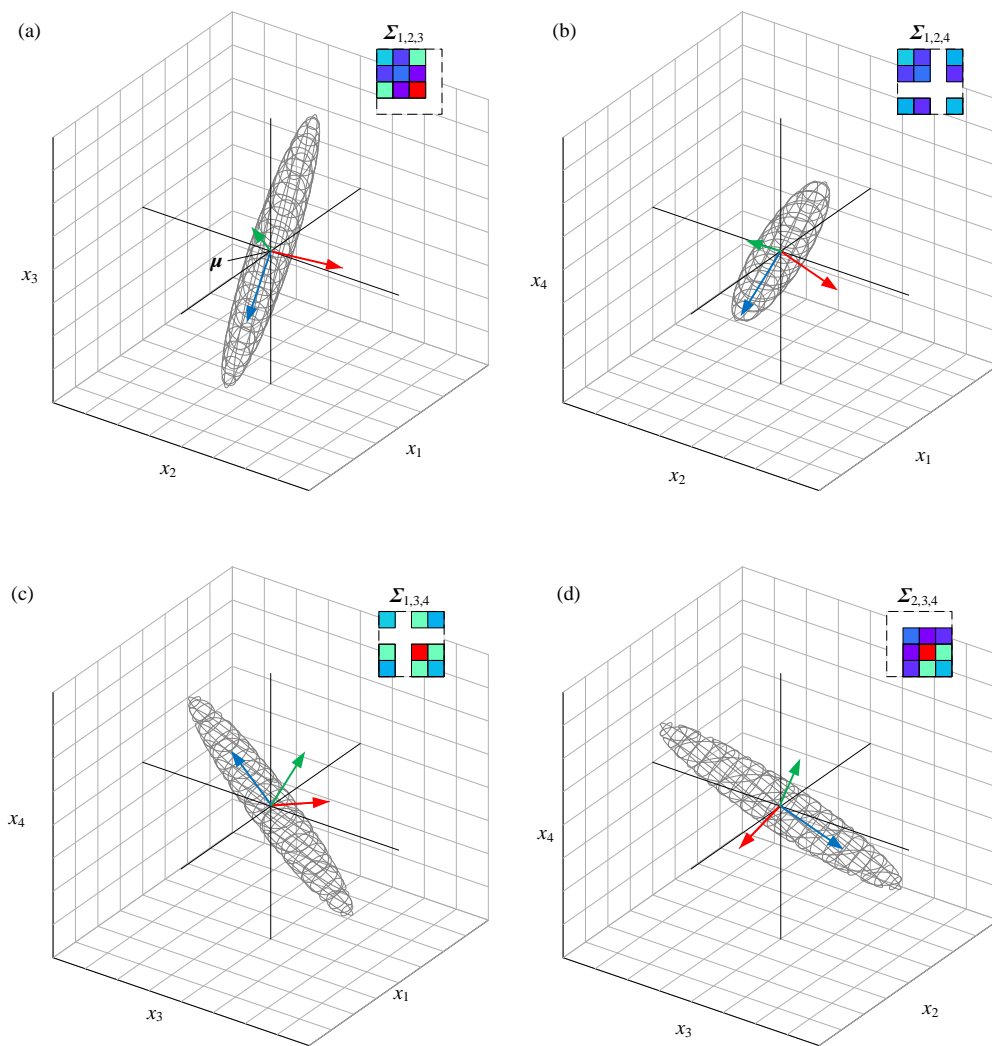


图 11 四维空间的“旋转”超椭球在三维空间中的四个投影

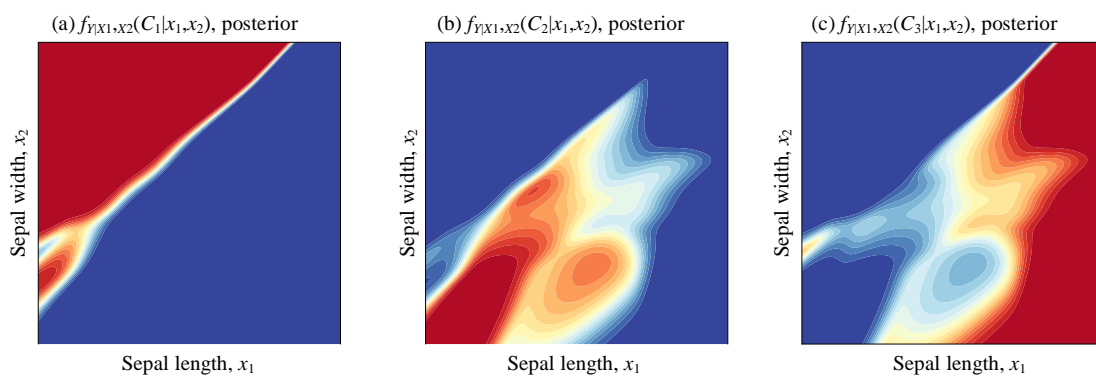


图 12. 比较三个后验概率曲面平面填充等高线

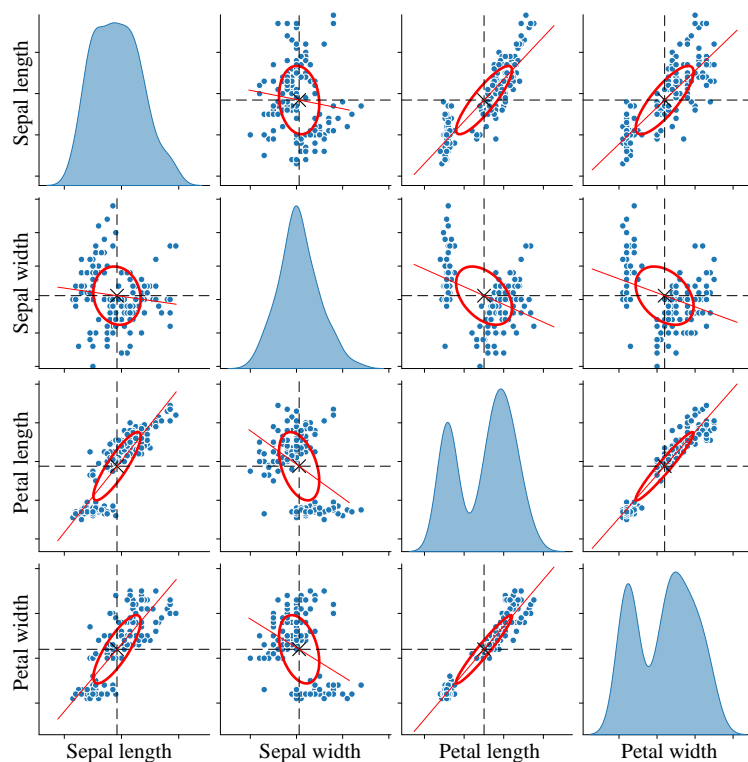
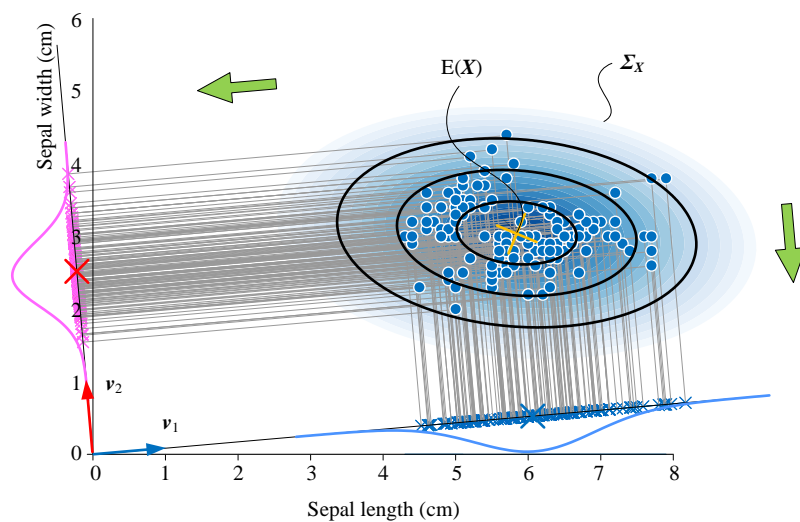


图 13. 鸢尾花数据成对特征图和回归关系

图 14. 鸢尾花数据正交系 V 投影

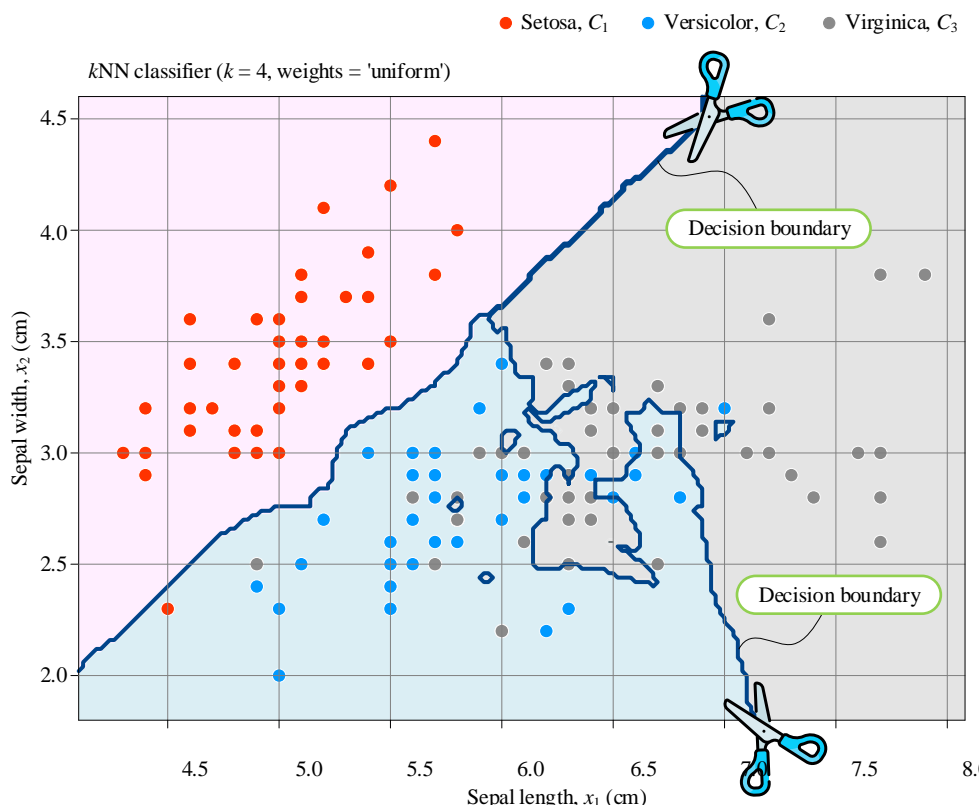


图 15. k 近邻分类, $k = 4$, 采用 2 个特征 (花萼长度 x_1 , 和花萼宽度 x_2) 分类三种鸢尾花

2.4 编程实现：Python 有大作为

《可视之美》之所以选择 Python 作为编程语言，是因为 Python 提供的是“一站式”解决方案。也就是说，从数据处理、数学运算，到统计分析、机器学习，乃至本书关注的可视化方案，Python 可谓应有尽有、一网打尽。

首先要强调的是，本书每一次使用 Python 编程完成可视化过程，都离不开数学工具；换个角度来看，创造各种不同可视化方案来分析同一个数学问题，或同一组数据的过程，都是一次次智力挑战。在这个过程中，我们不但画出了更有创意的图像，更好地掌握编程技巧；更重要的是，提升了自己的几何思维、数学思维。


请相信，经过你自己手写代码可视化的数学工具，这个数学工具的图像大概率这辈子应该就刻在你脑子里，怕是抹不去了。当大家实践的越多，见识的数学、编程工具越多，掌握的可视化方案越丰富，就越可能创造出更多更富创意的图像。

Python 拥有多种用于数据可视化的工具。以下是鸢尾花书中一些常用的可视化工具：

Matplotlib 是 Python 中最流行的绘图库之一，提供了广泛的绘图功能，包括折线图、散点图、柱状图、饼图等。它具有灵活性和广泛的定制选项，可以用于创建静态、交互式 and 动态的图形。本书大部分的静态矢量图都是用 Matplotlib 库函数完成；因此，大家大可以不用抱怨 Matplotlib 出图效果，我们需要的是提高自己的编程技能。一根根墨黑的碳条，也能绘出不朽的画作。

Seaborn 建立在 Matplotlib 之上的高级统计数据可视化库。Seaborn 提供了一组美观且具有统计意义的图表样式和绘图功能，使得数据的可视化变得更加简单和直观。

Plotly 是一个交互式的可视化库，支持多种图表类型，包括线图、散点图、柱状图、热力图等。Plotly 提供了丰富的交互功能，可以在网页中创建动态和可交互的图形。

图 18 ~ 图 20 提供了一个速查表，用来帮助大家找到合适的可视化方案以及对应的 Python 函数。表中， 代表该可视化函数具有交互属性。

除了以上提到的工具，还有其他一些流行的 Python 可视化库，如 Bokeh、Altair、ProPlot、Plotnine 等等，它们都提供了不同的特点和功能，可以根据具体需求选择适合的工具来进行数据可视化。

本书第 5 章将介绍 ProPlot，因为 ProPlot 出图效果特别类似《自然》等科技期刊。

2.5 美化完善：优化默认效果

利用 Python 编程可视化时，利用各种设置美化完善图像是重要的一环。

读过《编程不难》这本书的读者对图 16 都不陌生。

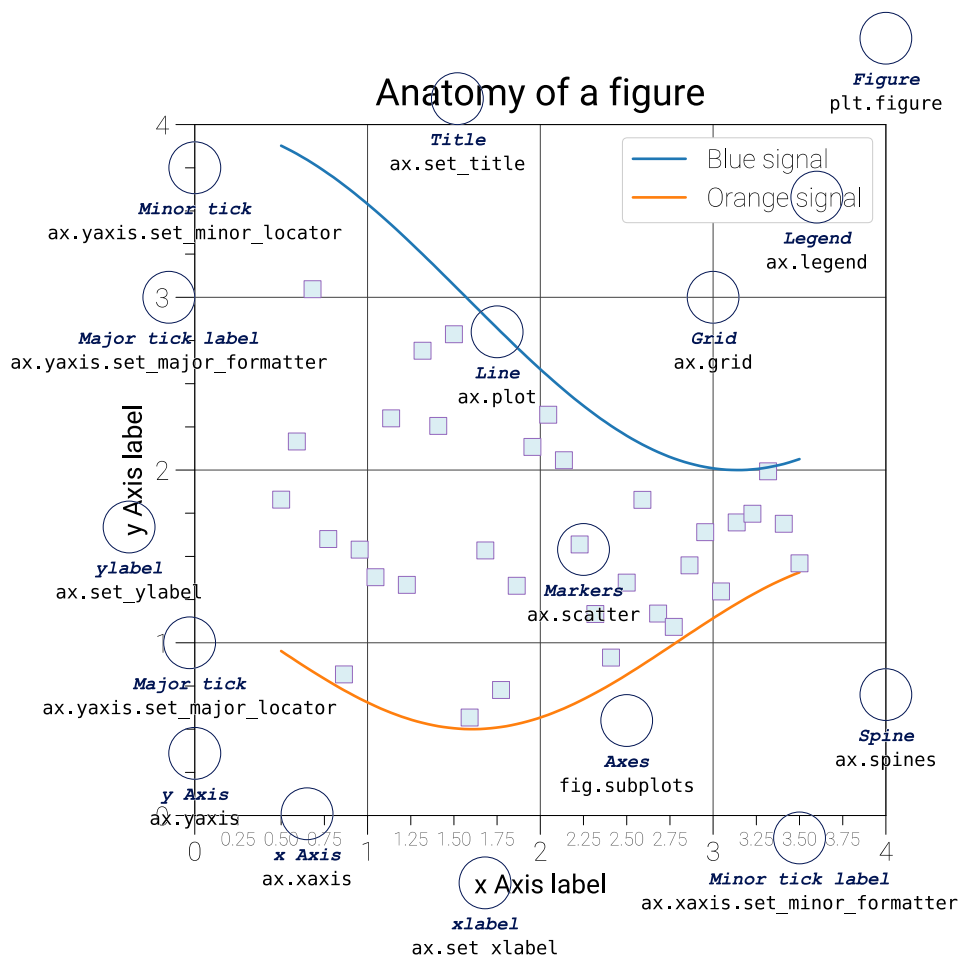


图 16. 解剖一幅图，来源 <https://matplotlib.org/stable/gallery/showcase/anatomy.html>

如图 16 所示，一幅图的基本构成部分包括以下几个部分：

- ▶ **图片对象 (figure)**：整个绘图区域的边界框，可以包含一个或多个子图。
- ▶ **子图对象 (axes)**：实际绘图区域，包含若干坐标轴、绘制的图像和文本标签等。
- ▶ **坐标轴 (axis)**：显示子图数据范围并提供刻度标记和标签的对象。
- ▶ **图脊 (spine)**：连接坐标轴和图像区域的线条，通常包括上下左右四条。
- ▶ **标题 (title)**：描述整个图像内容的文本标签，通常位于图像的中心位置或上方，用于简要概括图像的主题或内容。
- ▶ **刻度 (tick)**：刻度标记，表示坐标轴上的数据值。
- ▶ **标签 (label)**：用于描述坐标轴或图像的文本标签。
- ▶ **图例 (legend)**：标识不同数据系列的图例，通常用于区分不同数据系列或数据类型。
- ▶ **艺术家 (artist)**：在 Matplotlib 中，所有绘图元素都被视为艺术家对象，包括图像区域、子图区域、坐标轴、刻度、标签、图例等等。

美化完善时，以上组成部分都可以调整以便获得更好的可视化效果。本书下一版块专门介绍如何美化完善图像。

2.6 后期制作：丰富图片细节

后期制作是可视化重要环节之一。以鸢尾花书为例，用 Python 导出的矢量图不会被直接用到书稿中。每一幅都至少经过两个软件后期处理之后才会使用。常用的后期制作软件包括：Adobe Illustrator、Adobe Photoshop、Inkscape（免费）、Microsoft Visio 等。

为什么需要后期制作？下面给出几个理由：

- ▶ 美学设计：尽管 Python 库可以生成基本的图表，但在美学设计方面可能有一些限制。使用其他软件，如 Adobe Illustrator、Photoshop 或 Inkscape 等，可以提供更多的自定义选项，使图表更专业。
- ▶ 复杂效果：某些特殊效果可能在 Python 库中难以实现。其他软件通常提供更高级的图形处理功能。这些效果可以为图表增加视觉吸引力。
- ▶ 排版和注释：在生成图表后，可能需要添加额外的注释、标题、图例或其他文本元素；而编程增加这些元素可能耗时耗力，而且效果差、可编辑性差。其他软件通常提供更灵活的排版选项，可以更好地控制这些元素的位置和外观。
- ▶ 合并图表：如果需要将多个图表组合在一起，或者将图表与其他图像、照片或文本元素进行合并，其他软件通常提供更强大的合并和布局功能。

总之，尽管 Python 库提供了很多绘图功能，但有时候使用其他软件进行后期制作可以提供更大的自由度和更复杂的效果，以满足特定的需求。

注意，美化修饰没有问题，篡改数据必须坚决杜绝。

2.7 发布传播：到什么山上唱什么歌

经过了解规则、头脑风暴、编程实现、编程美化、后期制作等环节，一幅图算是诞生了，现在就差一步——发布传播媒介。

媒介是指传达信息和知识的工具或方式，它们以形式和技术将信息传递给受众。大家如果现在是通过正式出版图书阅读这段内容，那么媒介就是纸质。如果大家现在读的是 PDF 文件，那么媒介就是电子书。

以下是一些常见的知识传播媒介：

- ▶ 纸质是传统的知识传播媒介，比如报刊杂志、论文、书籍等等。纸质以印刷的方式将文字和图像展示在纸张上。纸质书具有持久性和便携性的优点，可以在没有电力或网络连接的情况下阅读。
- ▶ 电子书是以电子形式存在的书籍，比如，PDF、EPDB、LaTeX 等等。电子书可以在电子设备电子终端上阅读。电子书可以通过互联网下载或通过数字媒体载体传递。
- ▶ 幻灯片通常用于演示和演讲，可以包含文字、图像、图表和动画等内容。
- ▶ 网页是互联网上的文档，使用 HTML（超文本标记语言）编写。网页可以包含文本、图像、音频、视频和链接等多种元素，通过浏览器访问。
- ▶ 视频，比如 MP4、MOV、GIF 等等，是通过捕捉、录制和编辑连续图像帧来呈现动态场景的媒介。特别地，GIF（图形交换格式）支持动画和短视频片段。GIF 图像可以通过循环播放一系列图像帧来呈现动画效果。

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

- ▶ App 应用，提供丰富的交互功能，增强了用户的参与和学习体验。《编程不难》介绍过利用 Streamlit 制作 App 应用。

Python 中常见的绘图工具，比如 Matplotlib、Seaborn、Plotly，都可以生成各种格式的图片。这些图片可以用在纸质书、电子书、幻灯片、网页、视频等等各种媒体。

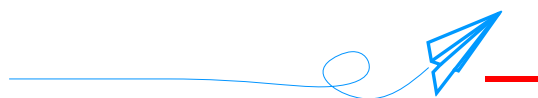
Matplotlib 可以通过 `matplotlib.pyplot.savefig()` 函数保存图片。《编程不难》介绍过 Matplotlib 导出的图片格式，下面简单盘点一下最常用的几种图片格式。

- ▶ SVG (Scalable Vector Graphics)，扩展名为 `.svg`，基于 XML 的矢量图格式，可以无损缩放和编辑。这也是鸢尾花书系列最常用的图片导出格式。
- ▶ PDF (Portable Document Format)，扩展名为 `.pdf`，是通用的跨平台文档格式，支持矢量图和位图，并能保留图形的高质量，适用于打印、展示和共享。大家如果要用 LaTeX 写文章，可能会用到这种图片格式。
- ▶ PNG (Portable Network Graphics)，扩展名为 `.png`，是无损的位图格式，支持透明度和高质量压缩。
- ▶ JPEG (Joint Photographic Experts Group)，扩展名为 `.jpg` 或 `.jpeg`，是常见的有损压缩格式，特别适用于存储和传输照片和复杂图像。注意，JPG/JPEG 不支持图像的透明度，这一点没有 PNG 方便。
- ▶ EPS (Encapsulated Postscript)，扩展名为 `.eps`，是基于 PostScript 语言的矢量图格式，支持高质量的打印输出，常用于出版和印刷领域。Adobe Illustrator 可以很轻松地处理 EPS 格式图形。

此外，Matplotlib 还可以生成 MP4、GIF 格式的视频文件。特别地，Plotly 支持交互图形可视化和 dashboard 搭建。

本章最后给大家提几个建议：

- ▶ 千万别抄袭！引用注明出处。
- ▶ 图片除了要美，还要有效，每一幅图都要服务于一条完整故事链。
- ▶ 控制时间成本。和数值相关的可视化部分，建议用编程实现；美化元素建议后期软件处理。
- ▶ 使用清晰的标签、标题和图例。轴标签、数值、单位、解说文字等尽量齐全。
- ▶ 某个可视化方案大量出图，建议写成通用函数。
- ▶ 风格尽量统一，比如颜色、线型、字体、字号、标注等等。
- ▶ 保持可视化的简洁明了。避免冗余装饰，确保重点和主要信息清晰可见。
- ▶ 将可视化放入适当的上下文中，以便更好地理解 and 解释数据和信息。提供相关背景信息、注释和说明，帮助观众理解可视化的含义和重要性。
- ▶ 选择适当的颜色和配色方案，以增强可视化的视觉吸引力和可读性。确保颜色的使用符合信息的含义，避免使用过多的颜色，以免造成混乱。
- ▶ 在电子媒体出版时，考虑为可视化添加交互性和动态效果，以增加用户参与和理解。例如，可以使用交互式工具让用户自由探索数据，或者使用动画效果展示变化和趋势。
- ▶ 倾听反馈，迭代升级。不断提高可视化技能。



艺术可以用来将数学工具、数据转化为生动的可视化方案，帮助人们更好地理解 and 解释数学原理、挖掘数据背后的故事。

在数学工具、创意编程、数据科学、机器学习、人工智能等应用场景，优质的“数学 + 艺术”可视化方案可以让人们发现数学之美、数据之美，甚至爱上数学。本书关注的正是这一点。

掌握编程可视化的过程可能会跌宕起伏，当然也会惊喜不断。与其抱怨工具不好用，不如把简单工具用好！

下面，正式邀请大家踏上本书的“数学 + 艺术”的美学动手实操之旅！

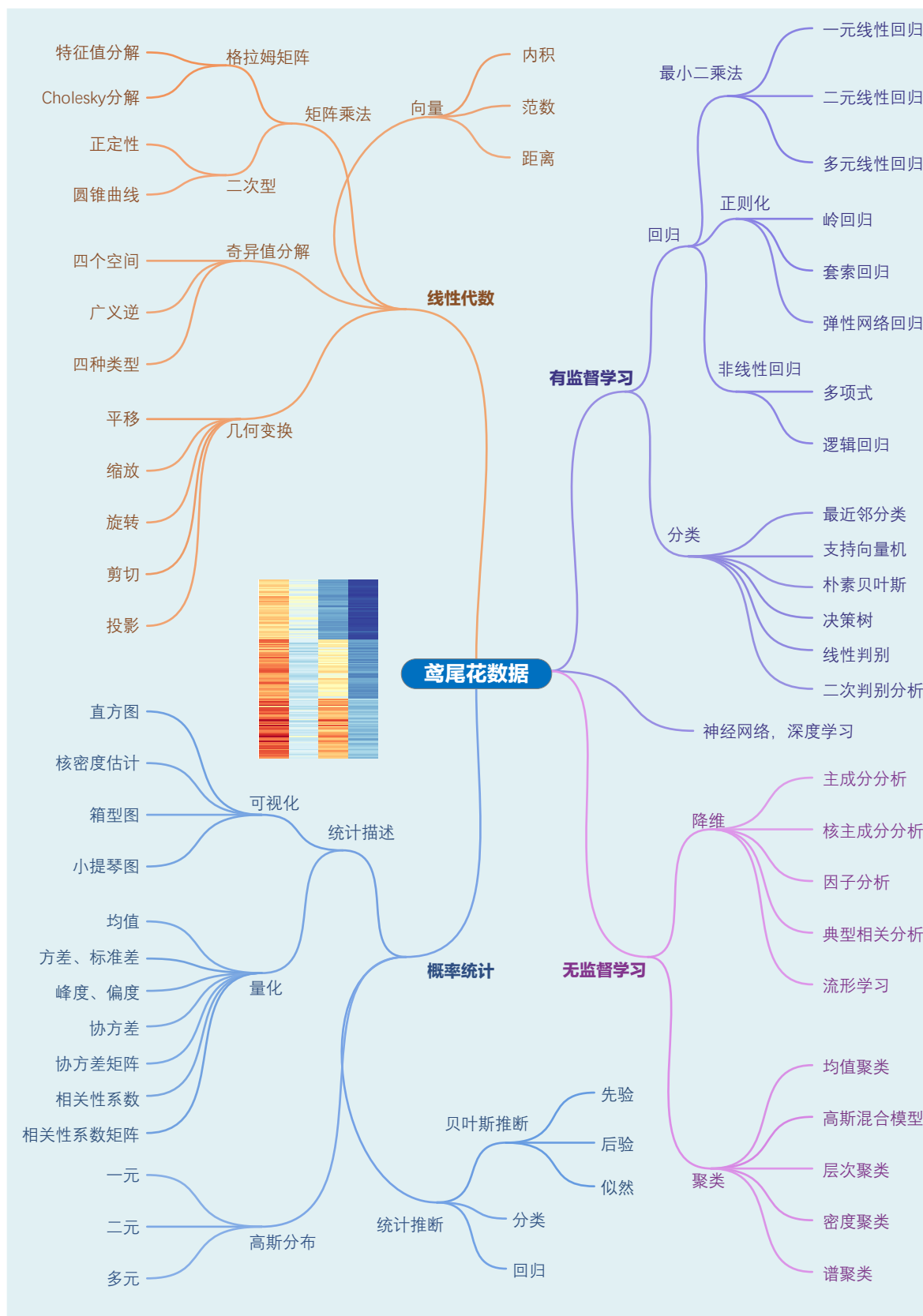


图 17. 有关鸢尾花数据的可视化“头脑风暴”

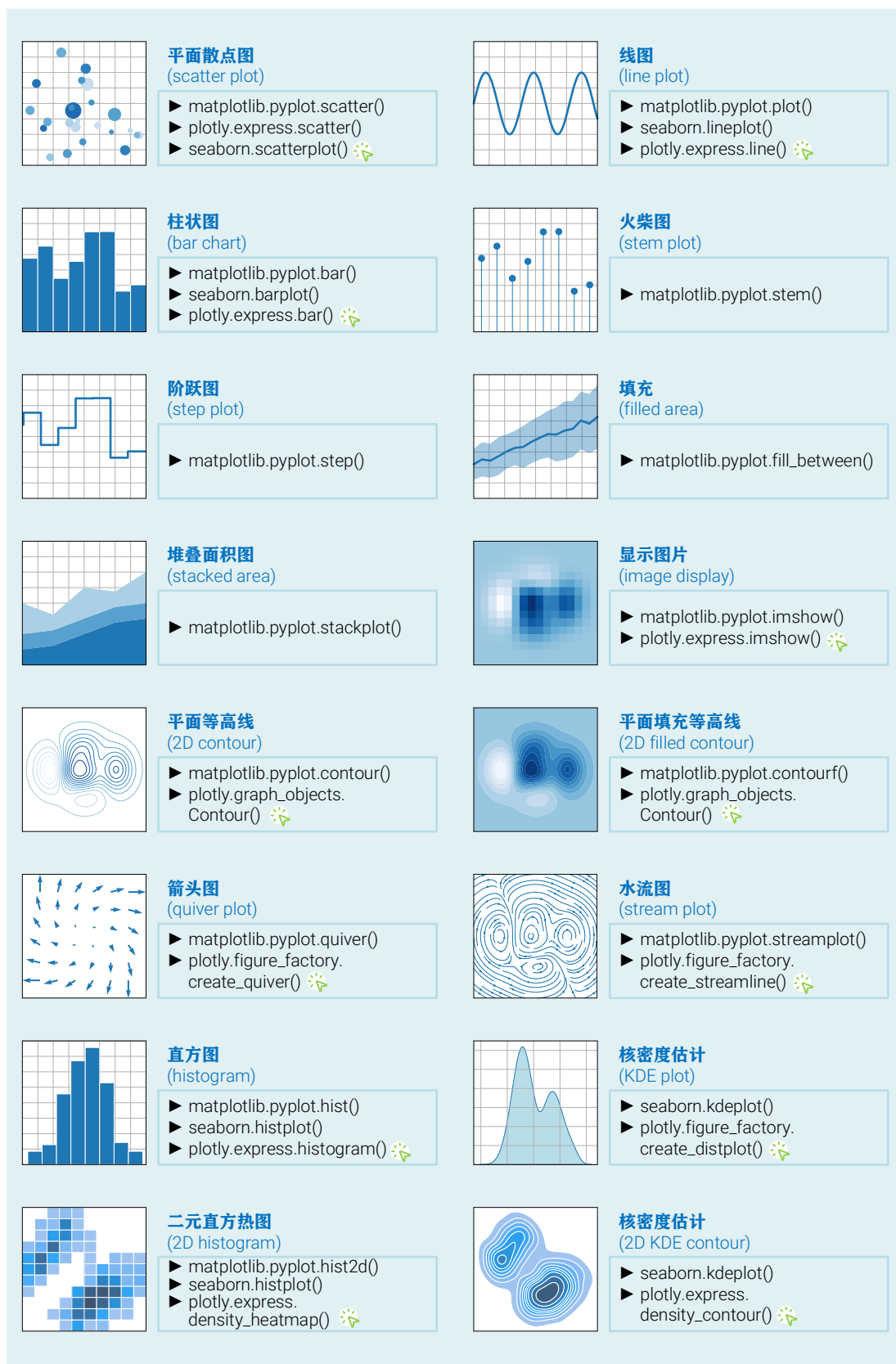


图 18. 鸢尾花书常用可视化方案目录，第 1 组

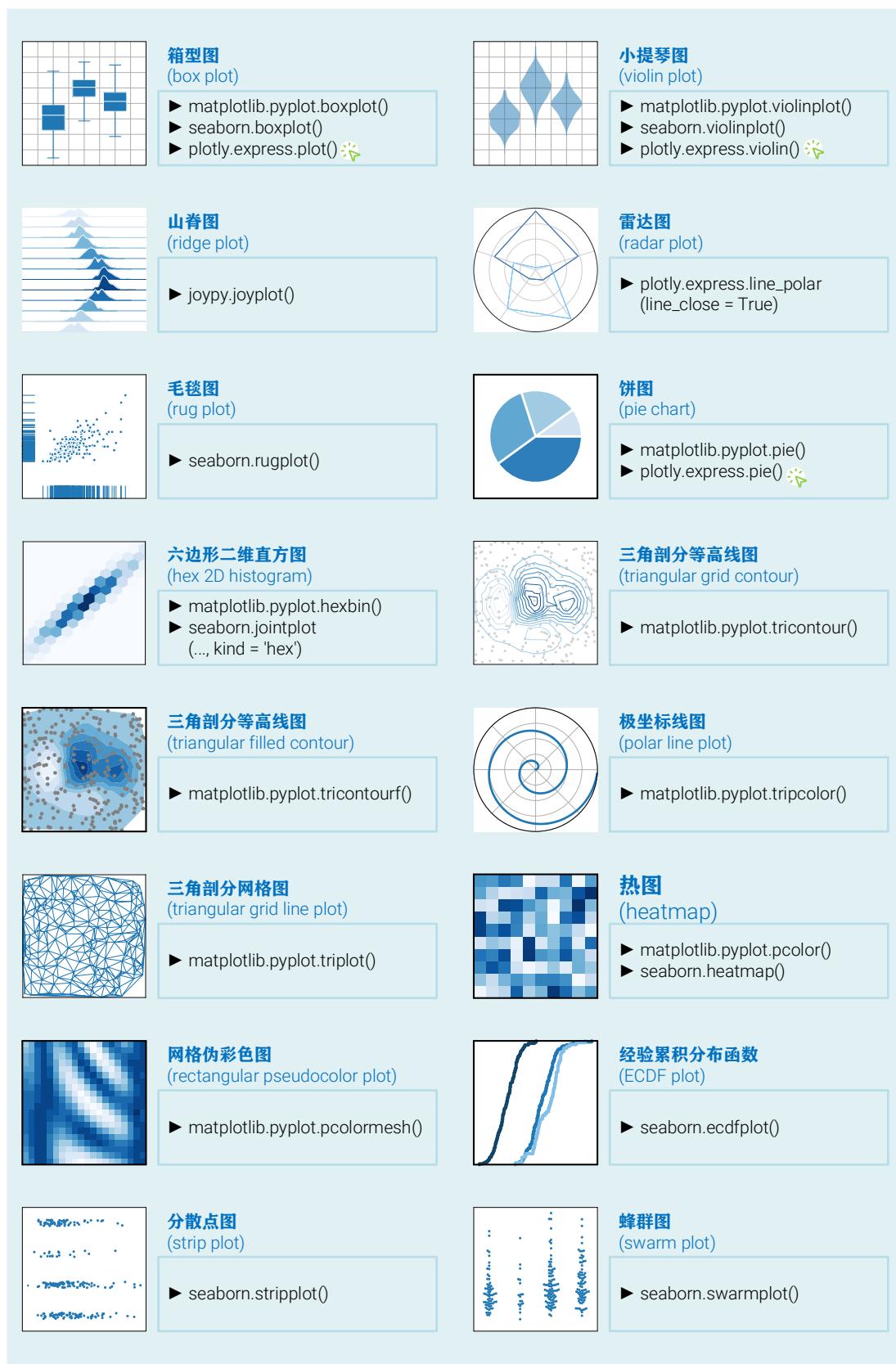


图 19. 鸢尾花书常用可视化方案目录, 第 2 组

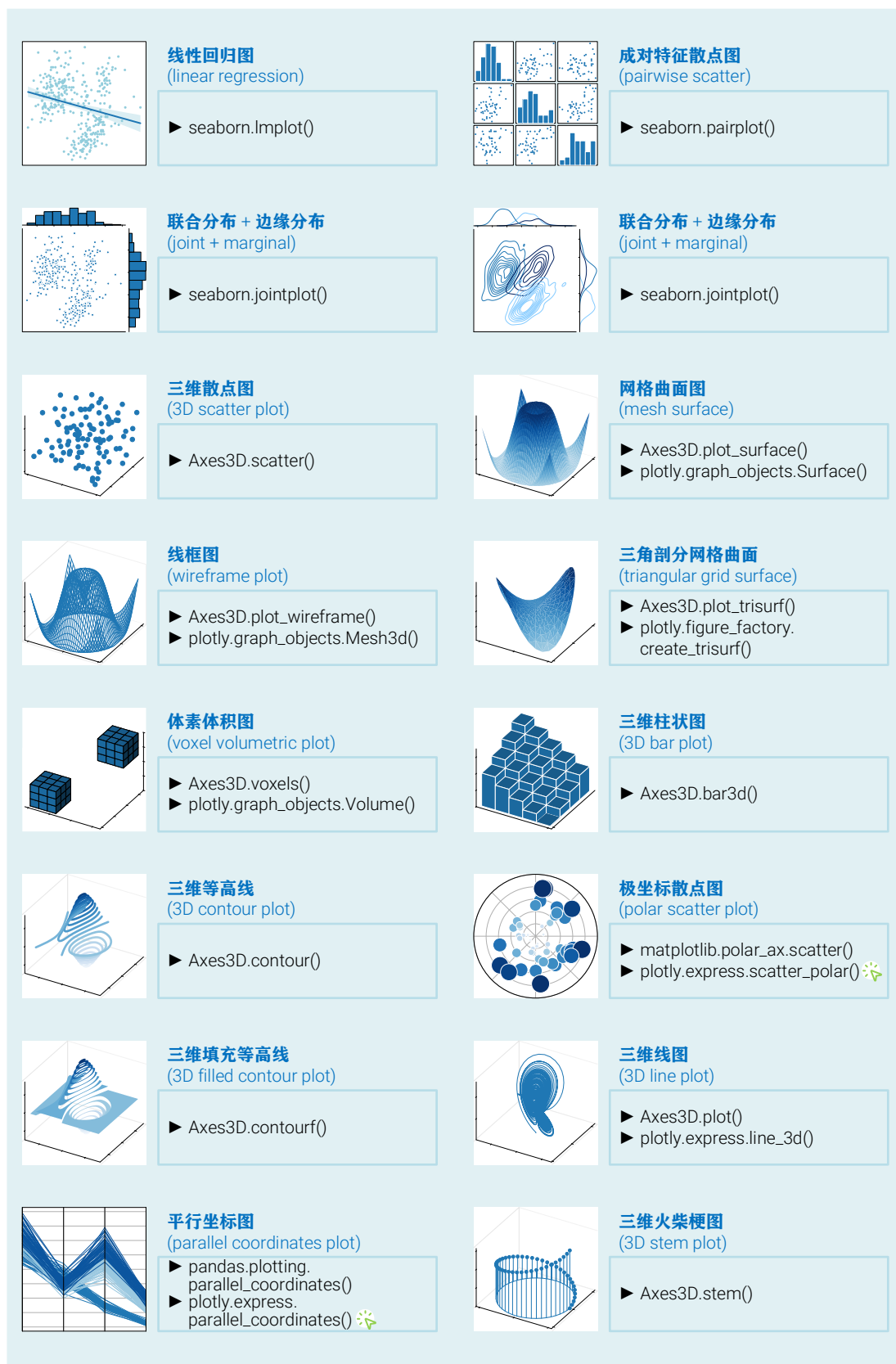


图 20. 鸢尾花书常用可视化方案目录, 第 3 组