

1

Vector and More

不止向量

一个有关向量的故事，从鸢尾花数据讲起



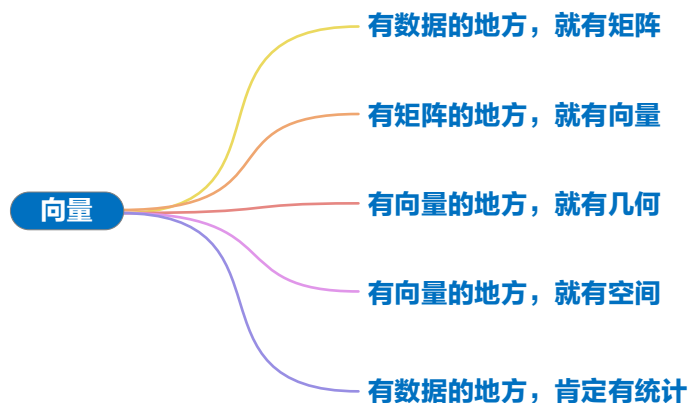
科学的每一次巨大进步，都源于颠覆性的大胆想象。

Every great advance in science has issued from a new audacity of imagination.

—— 约翰·杜威 (John Dewey) | 美国著名哲学家、教育家、心理学家 | 1859 ~ 1952



```
sklearn.datasets.load_iris() 加载鸢尾花数据
seaborn.heatmap() 绘制热图
```



1.1 有数据的地方，就有矩阵

本章主角虽然是**向量** (vector)，但是这个有关向量的故事先从**矩阵** (matrix) 讲起。

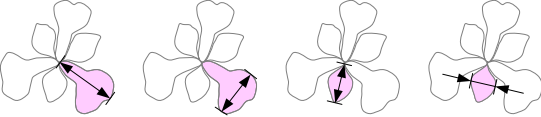
简单来说，矩阵是由若干行或若干列元素排列得到的**数组** (array)。矩阵内的元素可以是实数、虚数、符号，甚至是代数式。

从数据角度来看，矩阵就是表格！

数据科学、机器学习算法和模型都是“数据驱动”。没有数据，任何的算法都玩不转，数据是各种算法的绝对核心。“Garbage in, garbage out”。反之，优质数据本身就极具价值，不需要高深的算法分析数据，甚至不需要借助任何模型。

鸢尾花数据集

本书使用频率最高的数据是鸢尾花卉数据集。数据集的全称为**安德森鸢尾花卉数据集** (Anderson's Iris data set)，是植物学家**埃德加·安德森** (Edgar Anderson) 在加拿大魁北克加斯帕半岛上的采集的鸢尾花样本数据。图 1 所示为鸢尾花数据集部分数据。



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 1. 鸢尾花数据，数值数据单位为厘米 (cm)

图 1 给出的这些数据都属于鸢尾属下的三个亚属，分别是**山鸢尾** (setosa)、**变色鸢尾** (versicolor) 和**维吉尼亚鸢尾** (virginica)。每一类鸢尾花收集了 50 条样本记录，共计 150 条。

四个特征被用作样本的定量分析，它们分别是**花萼长度** (sepal length)、**花萼宽度** (sepal width)、**花瓣长度** (petal length) 和**花瓣宽度** (petal width)。

▲ 注意，本书用大写、粗体、斜体字母代表矩阵，比如 \mathbf{X} 、 \mathbf{A} 、 $\mathbf{\Sigma}$ 、 \mathbf{A} 。特别地，本书用 \mathbf{X} 代表样本数据矩阵，用 $\mathbf{\Sigma}$ 代表方差协方差矩阵 (variance covariance matrix)。本书用小写、粗体、斜体字母代表向量，比如 \mathbf{x} 、 \mathbf{x}_1 、 $\mathbf{x}^{(1)}$ 、 \mathbf{v} 。

如图2所示，本书常用**热图** (heatmap) 可视化矩阵。不考虑鸢尾花分类标签，鸢尾花数据矩阵 \mathbf{X} 有 150 行、4 列，因此 \mathbf{X} 也常记做 $\mathbf{X}_{150 \times 4}$ 。

前文提到，矩阵可以视作由一系列行向量、列向量构造而成。反向来看，矩阵切丝、切片可以得到行向量、列向量。如图2所示， \mathbf{X} 任一行向量 ($\mathbf{x}^{(1)}$ 、 $\mathbf{x}^{(2)}$ 、...、 $\mathbf{x}^{(150)}$) 代表一朵鸢尾花样本花萼长度、花萼宽度、花瓣长度和花瓣宽度测量结果， \mathbf{X} 某一列向量 (\mathbf{x}_1 、 \mathbf{x}_2 、 \mathbf{x}_3 、 \mathbf{x}_4) 为鸢尾花某个特征的样本数据。

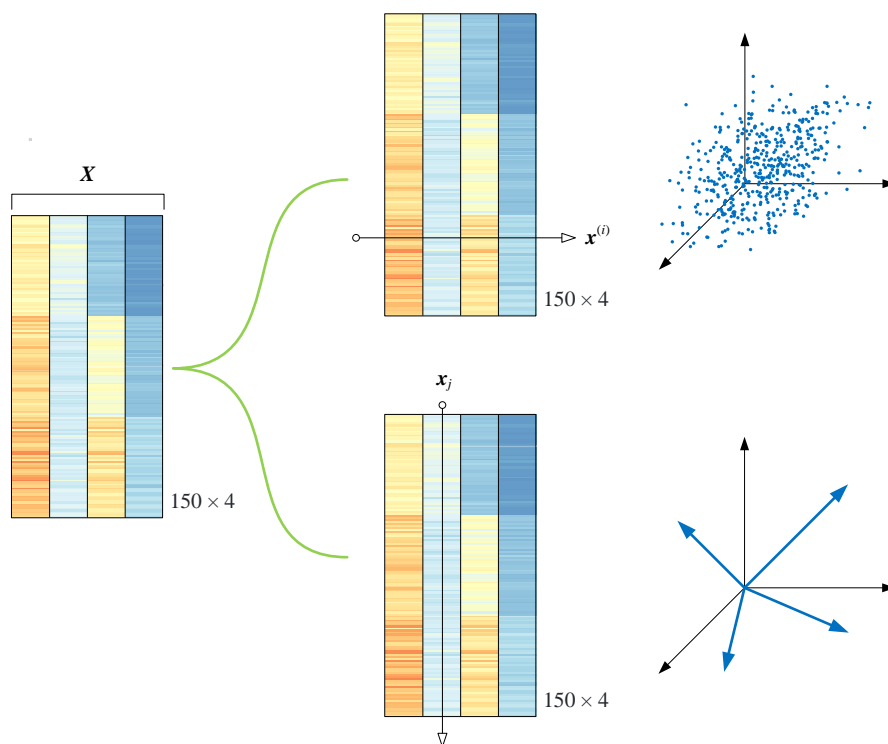


图2. 矩阵可以分割成一系列行向量或列向量

图片

数据矩阵其实无处不在。

再举个例子，大家日常随手拍摄的照片实际上就是数据矩阵。图3为作者拍摄的一张鸢尾花照片。把这张照片做黑白处理后，它变成了形状为 2990×2714 的矩阵，即 2990 行、2714 列。

图3这张照片显然不是矢量图。不断放大，我们会发现照片的局部变得越来越模糊。继续放大，我们发现这张照片竟然是由一系列灰度热图构成。再进一步，提取其中图片的4个像素点，也就是矩阵的4个元素，我们得到一个 2×2 实数矩阵。

➡ 本系列丛书《数据科学》将采用主成分分析 (Principal Component Analysis, PCA) 继续深入分析图3这幅鸢尾花黑白照片。

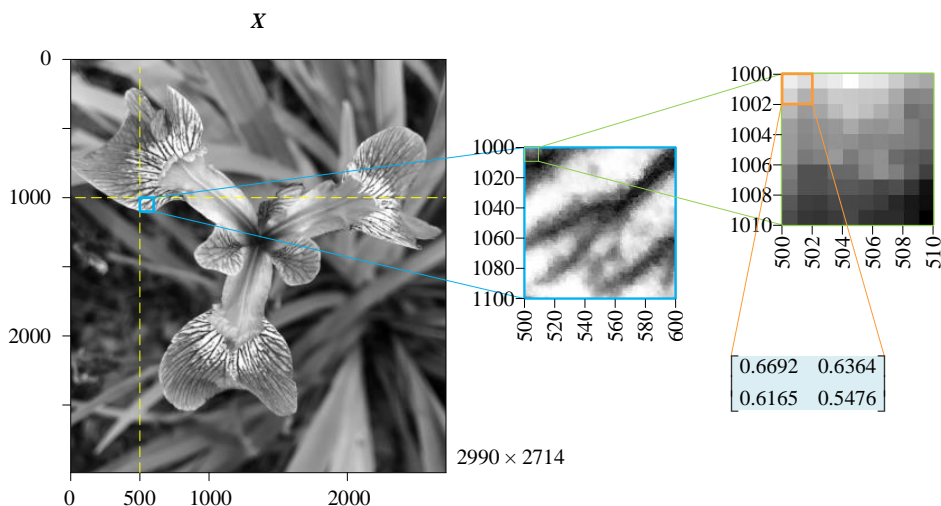


图3. 照片也是数据矩阵

1.2 有矩阵的地方，就有向量

行向量

首先，矩阵 X 可以看做是由一系列行向量 (row vector) 上下叠加而成。

如图4所示，矩阵 X 的第 i 行可以写成行向量 $\mathbf{x}^{(i)}$ 。上标圆括号中的 i 代表序号，对于鸢尾花数据集， $i = 1 \sim 150$ 。

举个例子， X 的第1行行向量记做 $\mathbf{x}^{(1)}$ ，具体为：

$$\mathbf{x}^{(1)} = [5.1 \quad 3.5 \quad 1.4 \quad 0.2]_{1 \times 4} \quad (1)$$

行向量 $\mathbf{x}^{(1)}$ 代表鸢尾花数据集编号为1的样本。行向量 $\mathbf{x}^{(1)}$ 的四个元素依次代表花萼长度 (sepal length)、花萼宽度 (sepal width)、花瓣长度 (petal length) 和花瓣宽度 (petal width)。长、宽度量单位均为厘米 cm。

行向量 $\mathbf{x}^{(1)}$ 也可以视作1行、4列的矩阵，即形状为 1×4 。

虽然 Python 是**基于 0 编号** (zero-based indexing)，本书对矩阵行、列编号时，还是延续线性代数传统，采用**基于 1 编号** (one-based indexing)。

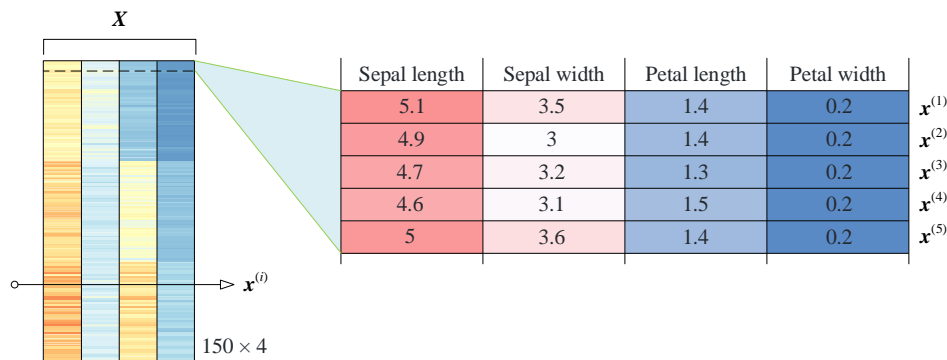


图 4. 鸢尾花数据，行向量代表样本数据点

列向量

矩阵 X 也可以视作一系列**列向量** (column vector) 左右排列而成。

如图 2 所示，矩阵 X 的第 j 列可以写成列向量 \mathbf{x}_j 。下标 j 代表列序号，对于鸢尾花数据集，不考虑分类标签的话， $j = 1 \sim 4$ 。

比如， X 的第 1 列向量记做 \mathbf{x}_1 ，具体为：

$$\mathbf{x}_1 = \begin{bmatrix} 5.1 \\ 4.9 \\ \vdots \\ 5.9 \end{bmatrix}_{150 \times 1} \quad (2)$$

列向量 \mathbf{x}_1 代表鸢尾花数据 150 个样本花萼长度数值。列向量 \mathbf{x}_1 可以视作 150 行、1 列的矩阵，即形状为 150×1 。整个数据矩阵 X 可以写成 $X = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4]$ 。

⚠ 注意，为了区分行向量和列向量，本书在编号时，行向量采用上标加圆括号，比如 $\mathbf{x}^{(1)}$ 。而列向量编号采用下标，比如 \mathbf{x}_1 。

➡ 大家可能会问， \mathbf{x}_1 、 \mathbf{x}_2 、 \mathbf{x}_3 、 \mathbf{x}_4 这四个向量到底意味着什么？有没有什么办法可视化这四个列向量？怎么量化它们之间的关系？答案会在本书第 12 章揭晓。

此外，大家熟悉的三原色光模式 (RGB color mode) 中每种颜色实际上也可以写成列向量，如图所示图 5 的 7 个颜色。在本书第 7 章中，我们将用 RGB 解释向量空间等概念。








						
$\begin{bmatrix} 0.8 \\ 0.8 \\ 0.8 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.8 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0.6 \\ 0.8 \\ 0.3 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0.7 \\ 0.9 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.8 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 0.3 \\ 0.3 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

图 5. 7 个颜色对应的 RGB 颜色向量

不要被向量、矩阵这些名词吓到。矩阵就是一个表格，而这个表格可以划分成若干行、若干列，它们分别叫行向量、列向量。

1.3 有向量的地方，就有几何

数据云、投影

取出鸢尾花前两个特征——花萼长度、花萼宽度——对应的数据。把它们以坐标的形式画在平面直角坐标系（记做 \mathbb{R}^2 ）中，我们便得到平面散点图。如图 6 所示，这幅散点图好比样本“数据云”。

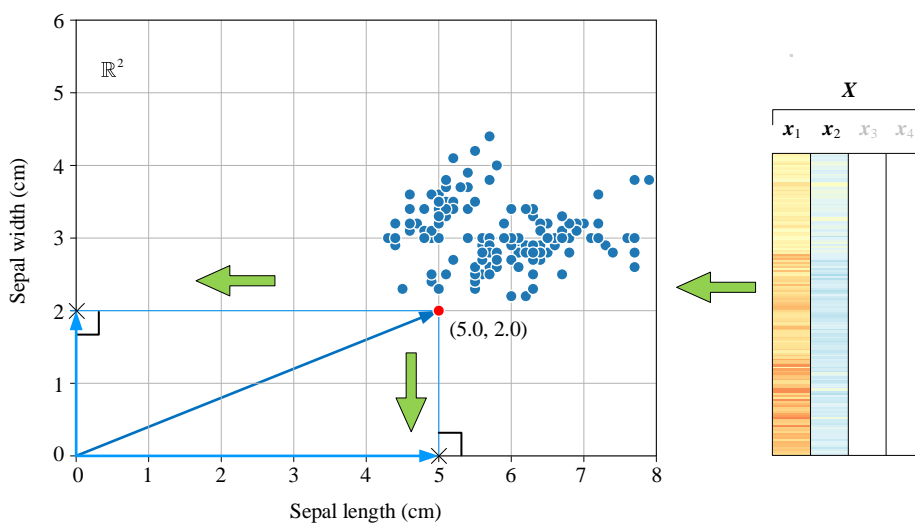


图 6. 鸢尾花前两个特征数据散点图

图 6 中数据点 $(5.0, 2.0)$ 可以写成行向量 $[5.0, 2.0]$ 。 $(5.0, 2.0)$ 是序号为 61 的样本点，对应的行向量可以写成 $\mathbf{x}^{(61)}$ 。

从几何视角来看， $[5.0, 2.0]$ 在横轴的**正交投影** (orthogonal projection) 结果为 5.0，代表它的横坐标为 5.0。 $[5.0, 2.0]$ 在纵轴的正交投影结果为 2.0，代表其纵坐标为 2.0。

正交 (orthogonality) 是线性代数的概念，是垂直的推广。正交投影很好理解，即原数据点和投影点连线垂直于投影点所在直线或平面。打个比方，头顶正上方阳光将自己身体的影子投影在地面，阳光光线垂直于地面。不特别强调的话，本书的投影均指正交投影。

从集合视角来看， $(5.0, 2.0)$ 属于平面 \mathbb{R}^2 ，即 $(5.0, 2.0) \in \mathbb{R}^2$ 。图 6 中整团数据云都属于 \mathbb{R}^2 。再者，如图 6 所示，从向量角度来看，行向量 $[5.0, 2.0]$ 在横轴上投影的向量为 $[5.0, 0]$ ，在纵轴上投影的向量为 $[0, 2.0]$ 。而 $[5.0, 0]$ 和 $[0, 2.0]$ 两个向量合成就是 $[5.0, 2.0]$ 。

再进一步，将图 6 整团数据云全部正交投影到横轴，得到图 7。图 7 中 \times 代表的数据实际上就是鸢尾花数据集第一列花萼长度数据。图 7 中横轴相当于一个一维空间，即数轴 \mathbb{R} 。

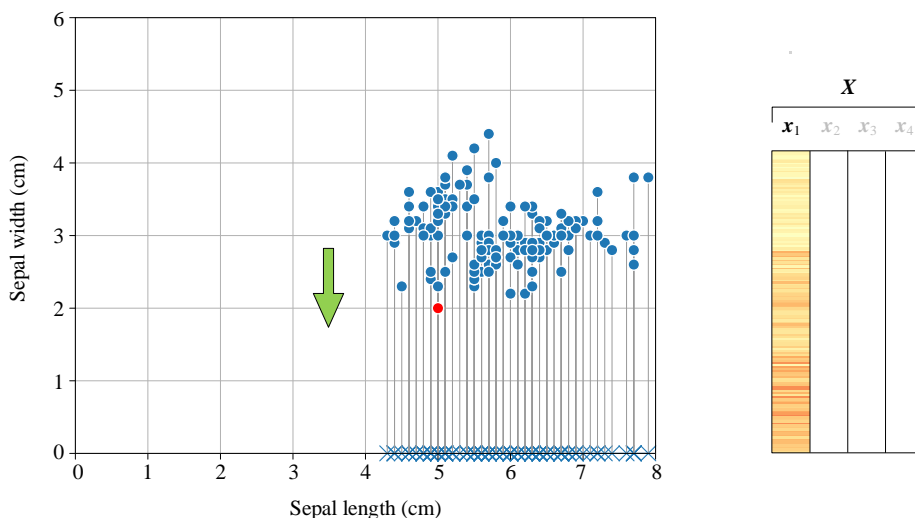


图 7. 二维散点正交投影到横轴

我们也可以把整团数据云全部投影在纵轴，得到图 8。图中的 \times 是鸢尾花数据集第二列花萼宽度数据。

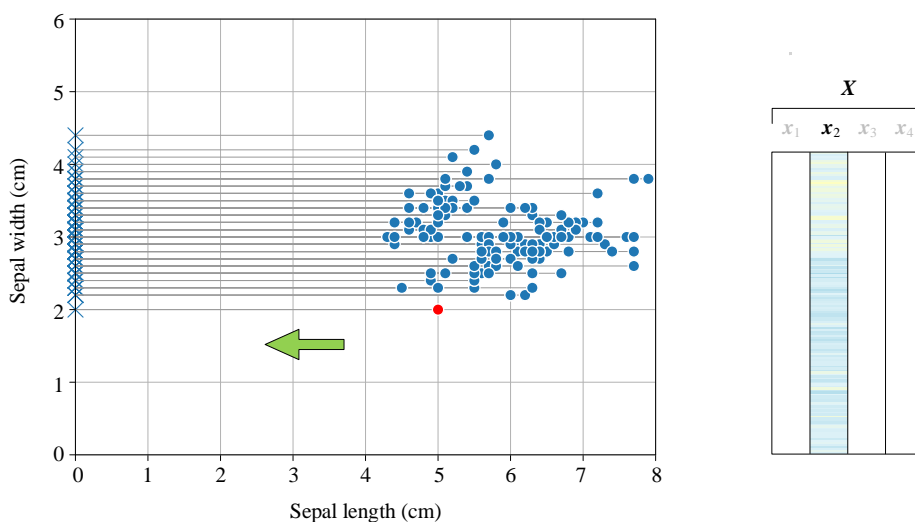


图 8. 二维散点正交投影到纵轴

你可能会问，是否可以将图 7 中所有点投影在一条斜线上？

答案是肯定的。

如图 9 所示，鸢尾花数据投影到一条斜线上，这条斜线通过原点和横轴夹角 15° 。观察图 9，我们已经发现投影点似乎是 x_1 和 x_2 的某种组合。也就是说， x_1 和 x_2 分别贡献 v_1x_1 和 v_2x_2 ，两种成分合成 $v_1x_1 + v_2x_2$ 就是投影点坐标。 $v_1x_1 + v_2x_2$ 也叫**线性组合** (linear combination)。

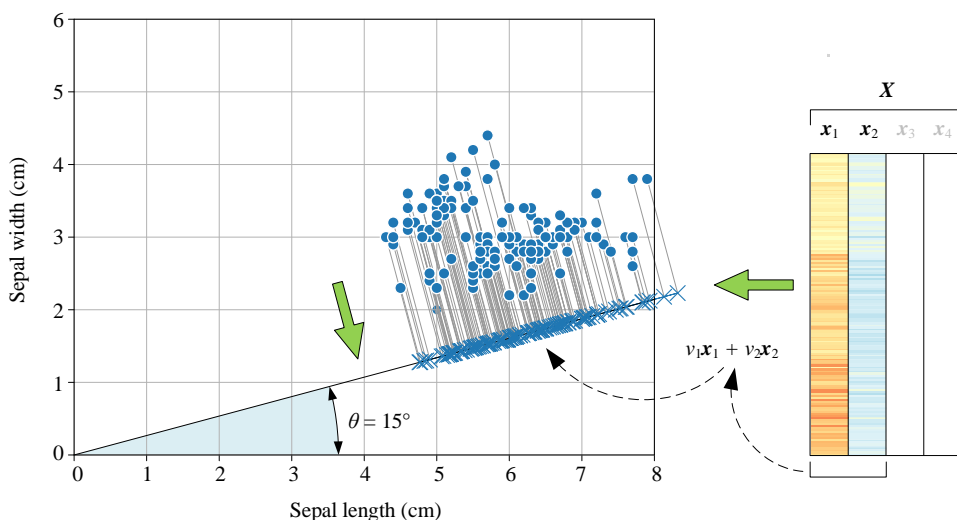


图 9. 二维散点正交投影到一条斜线

大家可能会问，怎么计算图 9 中投影点坐标？这种几何变换有何用途？这是本书第 9、10 章要回答的问题。

三维散点图、成对特征散点图

取出鸢尾花前三个特征（花萼长度、花萼宽度、花瓣长度）对应的数据，并在三维空间 \mathbb{R}^3 绘制散点图，得到图 10。图 6 相当于图 10 在水平面（浅蓝色背景）正交投影结果。

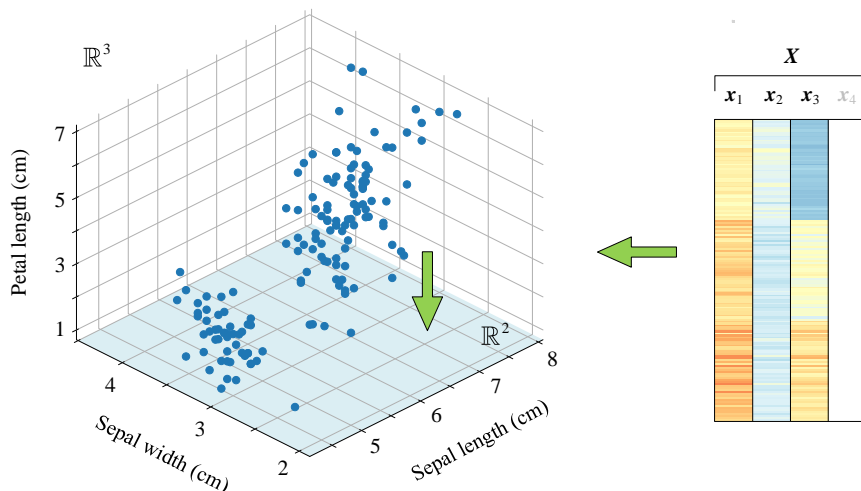


图 10. 鸢尾花前三个特征数据散点图

回顾本系列丛书《数学要素》一册介绍过的成对特征散点图，具体如图 11 所示。成对特征散点图不但可视化鸢尾花四个特征（花萼长度、花萼宽度、花瓣长度和花瓣宽度），通过散点颜色还可以展示鸢尾花三个类别（山鸢尾、变色鸢尾、维吉尼亚鸢尾）。图 11 中的每一幅散点图相当于四维空间数据在不同平面上的投影结果。

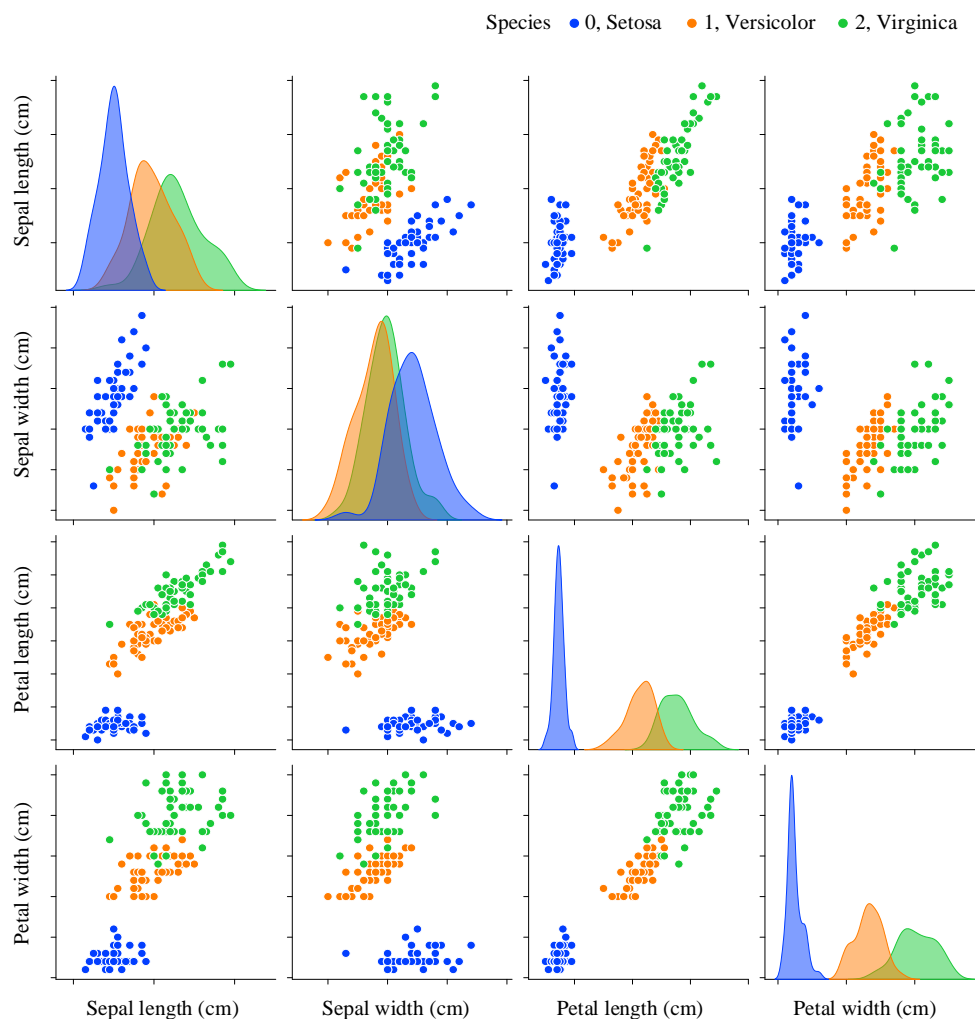


图 11. 鸢尾花数据成对特征散点图，考虑分类标签，图片来自《数学要素》

向量起点

如图 12 所示，本节前行向量的起点都是原点，即零向量 $\mathbf{0}$ 。但是，统计视角下，向量的起点移动到了数据**质心** (centroid)。所谓数据质心就是数据每一特征均值构成的向量。

如图 13 所示，将向量的起点移动到质心后，向量的长度、绝对角度（比如，和坐标系横轴夹角）、相对角度（向量两两之间的夹角）都发生了显著变化。

这一点也不难理解，大家回想一下，我们在计算方差、均方差、协方差、相关性系数等统计度量时，都会去均值。从向量角度来看，这相当于移动向量起点。

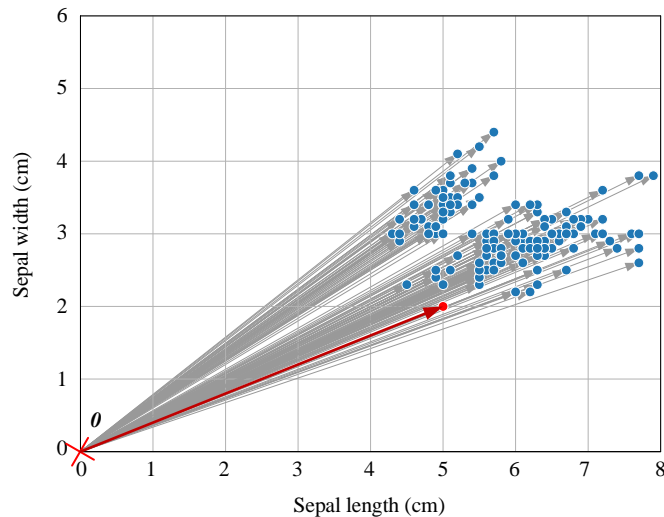


图 12. 向量起点为原点

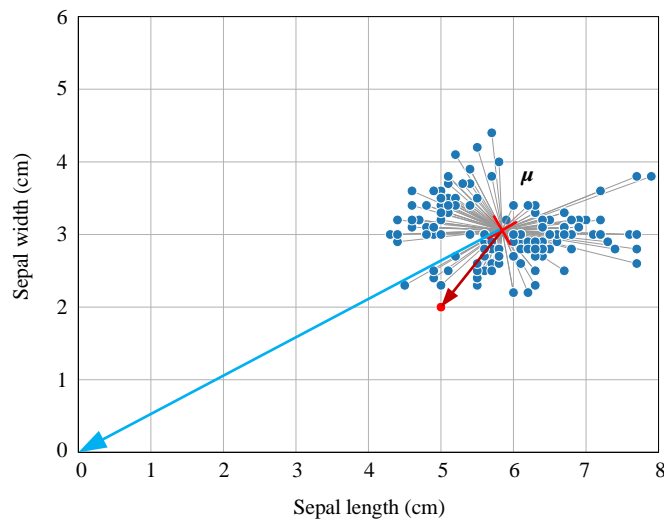


图 13. 向量起点为质心

将图 13 整团数据云质心平移到原点，这个过程就是去均值过程，结果如图 14 所示。数据矩阵 \mathbf{X} 去均值化得到的数据矩阵记做 \mathbf{X}_c ，显然 \mathbf{X}_c 的质心位于原点 $\mathbf{0}$ 。



观察图 11，我们发现，如果考虑数据标签的话，每一类标签样本数据都有自己质心，叫做分类质心，这是本书第 22 章要讨论的话题。此外，本书最后三章——数据三步曲——会把数据、矩阵、向量、矩阵分解、空间、优化、统计等板块联结起来。

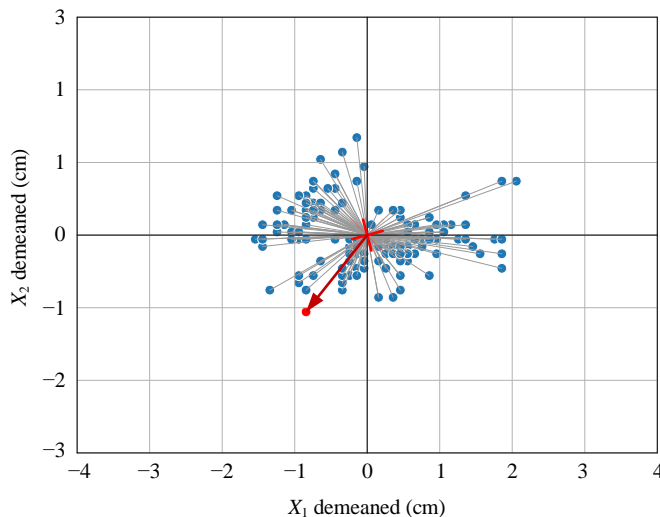


图 14. 数据去均值化

1.4 有向量的地方，就有空间

从线性方程组说起

从代数视角来看，**矩阵乘法** (matrix multiplication) 代表**线性映射** (linear mapping)。比如，在 $\mathbf{A}_{m \times n} \mathbf{x}_{n \times 1} = \mathbf{b}_{m \times 1}$ 中矩阵 $\mathbf{A}_{m \times n}$ 扮演的角色就是 $\mathbb{R}^n \rightarrow \mathbb{R}^m$ 线性映射。

在本系列丛书《数学要素》“鸡兔同笼三部曲”中，我们用线性方程组解决过鸡兔同笼问题。下面简单回顾一下。

《孙子算经》这样引出鸡兔同笼问题：“今有雉兔同笼，上有三十五头，下有九十四足，问雉兔各几何？”

将这个问题写成**线性方程组** (system of linear equations)：

$$\begin{cases} 1 \cdot x_1 + 1 \cdot x_2 = 35 \\ 2 \cdot x_1 + 4 \cdot x_2 = 94 \end{cases} \Rightarrow \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\mathbf{x}} = \underbrace{\begin{bmatrix} 35 \\ 94 \end{bmatrix}}_{\mathbf{b}} \quad (3)$$

即：

$$Ax = b \quad (4)$$

未知变量构成的列向量 x 可以利用下式求解：

$$x = A^{-1}b = \begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}^{-1} \begin{bmatrix} 35 \\ 94 \end{bmatrix} = \begin{bmatrix} 2 & -0.5 \\ -1 & 0.5 \end{bmatrix} \begin{bmatrix} 35 \\ 94 \end{bmatrix} = \begin{bmatrix} 23 \\ 12 \end{bmatrix} \quad (5)$$



(5) 用到了矩阵乘法 (matrix multiplication)、矩阵逆 (matrix inverse)。本书第 4、5、6 三章将介绍矩阵相关运算，居于核心的运算当属矩阵乘法。

几何视角

从几何视角来看，(3) 中矩阵 A 完成的是线性变换 (linear transformation)。如图 15 所示，矩阵 A 把 e_1 和 e_2 构成的方方正正的方格，变成平行四边形网格，对应的计算为：

$$\underbrace{\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}}_A \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{e_1} = \underbrace{\begin{bmatrix} 1 \\ 2 \end{bmatrix}}_{a_1}, \quad \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}}_A \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{e_2} = \underbrace{\begin{bmatrix} 1 \\ 4 \end{bmatrix}}_{a_2} \quad (6)$$

而上式结果恰好是矩阵 $A = [a_1, a_2]$ 的两个列向量 a_1 和 a_2 。

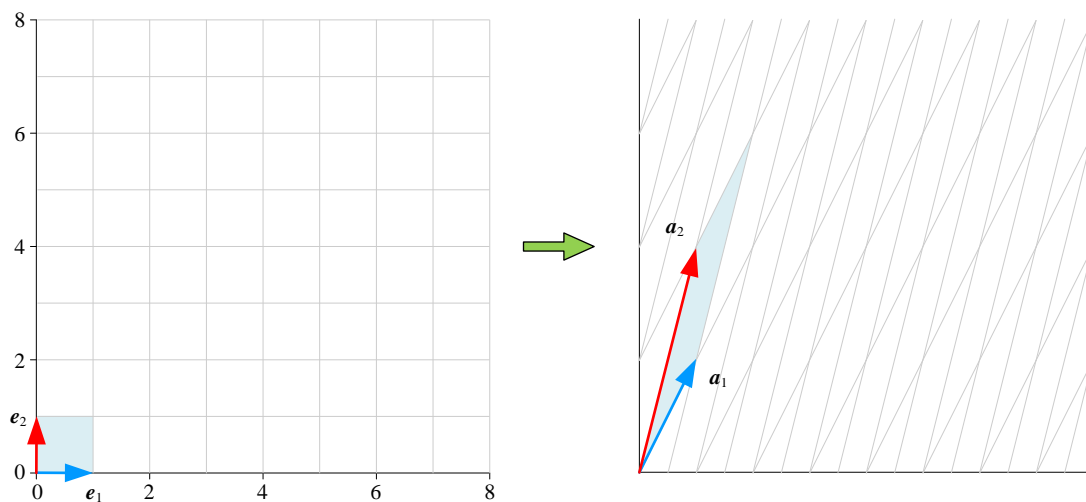


图 15. 矩阵 A 完成的线性变换

观察图 15 左图，整个直角坐标系整个方方正正的网格由 $[e_1, e_2]$ 张成，就好比 e_1 和 e_2 是撑起这个二维空间的“骨架”。再看图 15 右图， $[a_1, a_2]$ 同样张成了整个直角坐标系，不同的是网格为平行四边形。 $[e_1, e_2]$ 和 $[a_1, a_2]$ 都是各自空间的基底 (base)。

将 A 写成 $[a_1, a_2]$ ，展开 (4) 得到：

$$[a_1 \ a_2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 a_1 + x_2 a_2 = b \quad (7)$$

上式代表基底 $[a_1, a_2]$ 中两个基底向量的线性组合。



本书将在第 7 章专门讲解向量空间。

从正圆到旋转椭圆

椭圆等圆锥曲线在本系列丛书扮演重要角色，这一切都源于多元高斯分布概率密度函数。而线性变换和椭圆又有千丝万缕的联系。

如图 16 所示，同样利用 (3) 中矩阵 A ，我们可以把一个单位圆转化为旋转椭圆。图 16 中，终点落在单位圆上的任意向量 x ，经过 A 的线性变换变成 Ax 。

图 16 旋转椭圆的半长轴长度约为 4.67，半短轴长度约为 0.43，半短轴和横轴夹角约为 -16.85° 。要完成这些计算，我们需要线性代数中一个利器——**特征值分解** (eigen decomposition)。



本书读者对特征值分解并不陌生，如图 17 所示，我们在本系列丛书《数学要素》鸡兔同笼三部“鸡兔互变”中简单聊过特征值分解，大家如果忘记了，建议回顾一下。本书第 13、14 章专门探讨特征值分解。此外，本书将在第 20、21 章利用线性代数工具分析圆锥曲线和二次曲面。

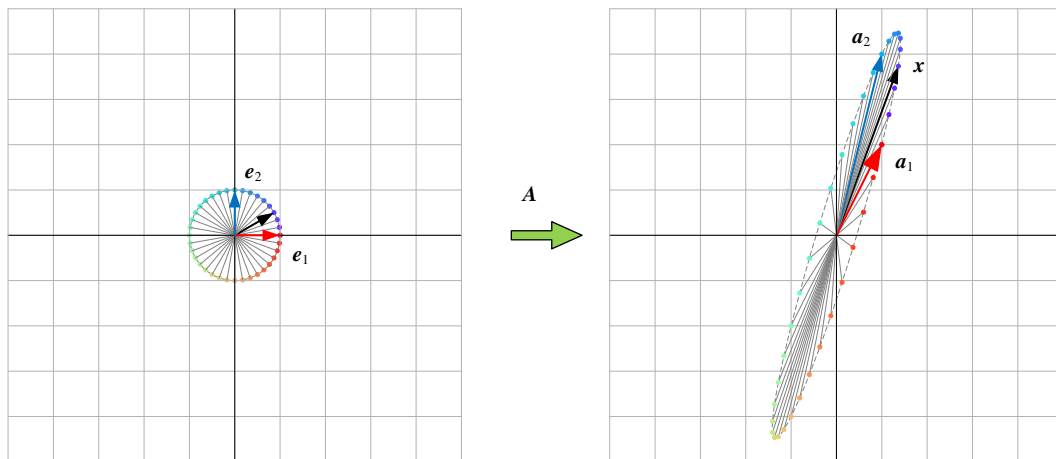


图 16. 矩阵 A 将单位圆转化为旋转椭圆

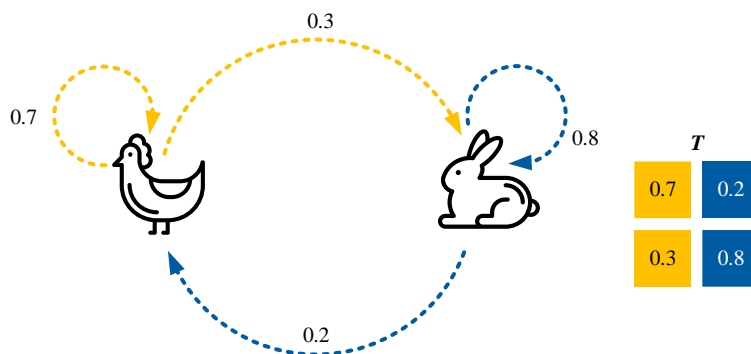
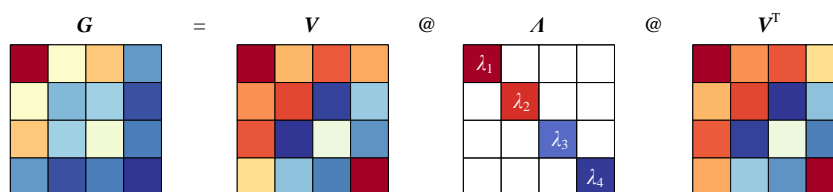


图 17. 鸡兔同笼三部曲中“鸡兔互变”，图片来自本系列丛书《数学要素》第 25 章

特征值分解

剧透一下，鸢尾花数据矩阵 X 本身并不能完成特征值分解。但是图 21 中的格拉姆矩阵 $G = (X^T X)$ 可以完成特征值分解，分解过程如图 18 所示。请大家特别注意图 18 中的矩阵 V 。正如图 15 右图中 $A = [a_1, a_2]$ 张成了一个平面，矩阵 $V = [v_1, v_2, v_3, v_4]$ 则张成了一个 4 维空间！

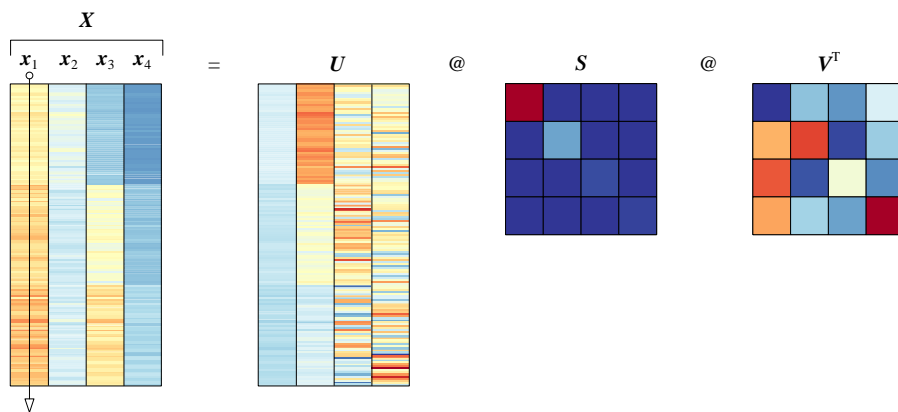
图 18. 矩阵 X 的格拉姆矩阵的特征值分解

奇异值分解

在**矩阵分解** (matrix decomposition) 这个工具库中，最全能的工具叫**奇异值分解** (Singular Value Decomposition, SVD)。图 19 所示为对鸢尾花数据矩阵的 SVD 分解，这幅图中的 U 和 V 都各自张成不同的空间。



本书第 15、16 章专门讲解奇异值分解，第 23 章则利用 SVD 分解引出四个空间。

图 19. 对矩阵 X 进行 SVD 分解

1.5 有数据的地方，肯定有统计

前文提到，图 20 所示鸢尾花数据每一列代表鸢尾花的一个特征，比如花萼长度（第 1 列，列向量 x_1 ）、花萼宽度（第 2 列，列向量 x_2 ）、花瓣长度（第 3 列，列向量 x_3 ）和花瓣宽度（第 4 列，列向量 x_4 ）。这些列向量可以看成是 X_1 、 X_2 、 X_3 、 X_4 四个随机变量的样本值集合。

从统计视角来看，我们可以计算样本数据各个特征的均值 (μ_j)，计算不同特征上样本数据的均方差 (σ_j)。图 20 中四副子图中的曲线代表各个特征样本数据的**概率密度估计** (probability density estimation) 曲线。有必要的話，我们还可以在图中标出 μ_j 、 $\mu_j \pm \sigma_j$ 、 $\mu_j \pm 2\sigma_j$ 对应的位置。

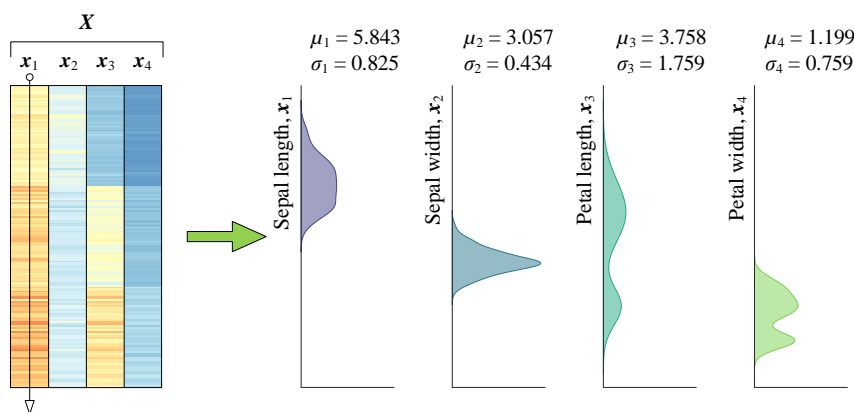


图 20. 鸢尾花数据，列向量代表数据特征

实际应用时，我们还会对原始数据进行处理，常见的操作有**去均值** (demean)、**标准化** (standardization) 等。

本系列丛书《数据科学》将专门介绍缺失值、离群值、数据转换、插值等数据处理问题。

对于多个特征之间的关系，我们可以采用**格拉姆矩阵** (Gram matrix)、**协方差矩阵** (covariance matrix)、**相关性系数矩阵** (correlation matrix) 等矩阵来描述。

图 21 所示为本书后续要用到的鸢尾花数据矩阵 X 衍生得到的几种矩阵。注意，图 2 和图 21 矩阵 X 热图采用不同的色谱值。

本书第 22 章将介绍如何获得图 21 所示这些矩阵，本书第 24 章将探讨图 21 主要矩阵和各种矩阵分解 (matrix decomposition) 之间有趣关系。

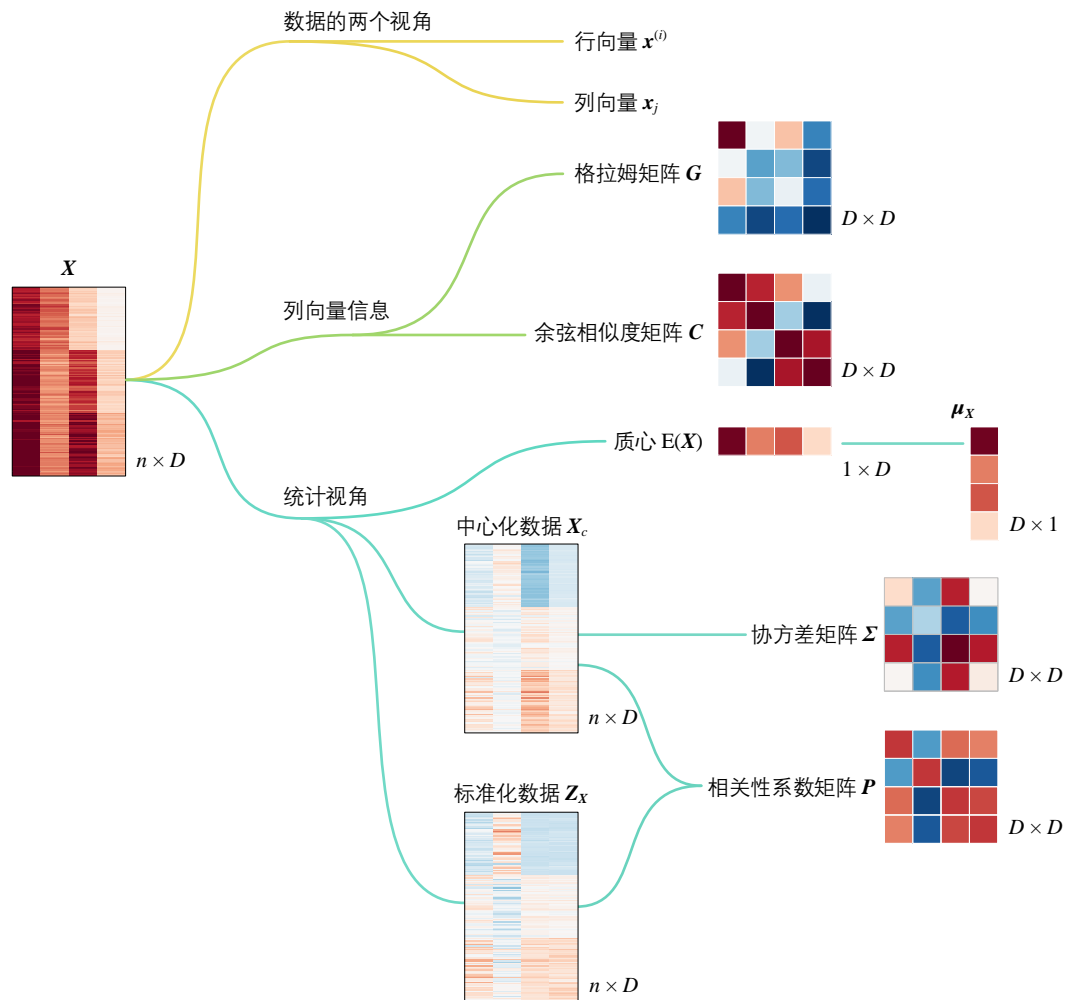


图 21. 鸢尾花数据衍生得到的几个矩阵



本章只配套一个代码文件，Streamlit_Bk4_Ch1_01.py。这段代码中，我们用 Streamlit 和 Plotly 分别绘制了鸢尾花数据集的热图、平面散点图、三维散点图、成对特征散点图。这四幅图都是可交互图像。



本章以向量为主线，回顾了《数学要素》“鸡兔同笼三部曲”的主要内容，预告了本书主要内容。不需要大家理解本章提到所有术语，只希望大家记住以下几句话：

有数据的地方，就有矩阵！

有矩阵的地方，就有向量！

有向量的地方，就有几何！

有向量的地方，就有空间！

有数据的地方，肯定有统计！



对线性代数概念感到困惑的读者，推荐大家看看 3Blue1Brown 制作的视频。很多视频网站上都可以找到译制视频。