

# 1

## Vector and More

# 不止向量

有关向量的故事，从鸢尾花数据讲起



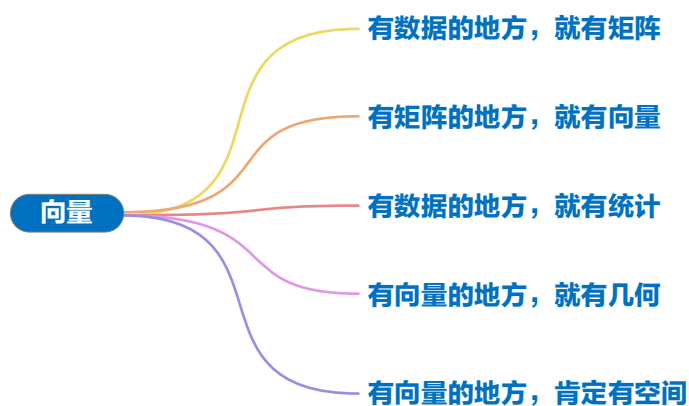
科学的每一次巨大进步，都源于颠覆性的大胆想象。

*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。

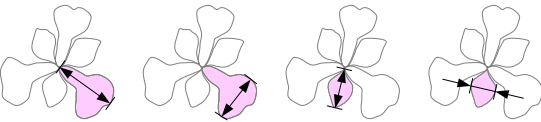
简单来说，矩阵是由若干行或若干列元素排列得到的**数组** (array)。矩阵内的元素可以是实数、虚数、符号，甚至是代数式。向量可以看做是一维数组，而矩阵是二维数组。

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 150 个鸢尾花样本数据。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)。

数据整体可以看做是一个矩阵  $\mathbf{X}$ ， $\mathbf{X}$  列向量为鸢尾花某个特征的样本数据， $\mathbf{X}$  的行向量代表一朵鸢尾花不同特征的数值。

▲ 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，我们用  $\mathbf{X}$  代表样本数据矩阵，用  $\mathbf{\Sigma}$  代表方差协方差矩阵。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  $\mathbf{X}_{150 \times 4}$ 。

矩阵可以视作由一系列行向量、列向量构造而成。反方向来看，矩阵切丝、切片可以得到行向量、列向量。如图 2 所示，不考虑最后一列分类标签， $\mathbf{X}$  每一行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。

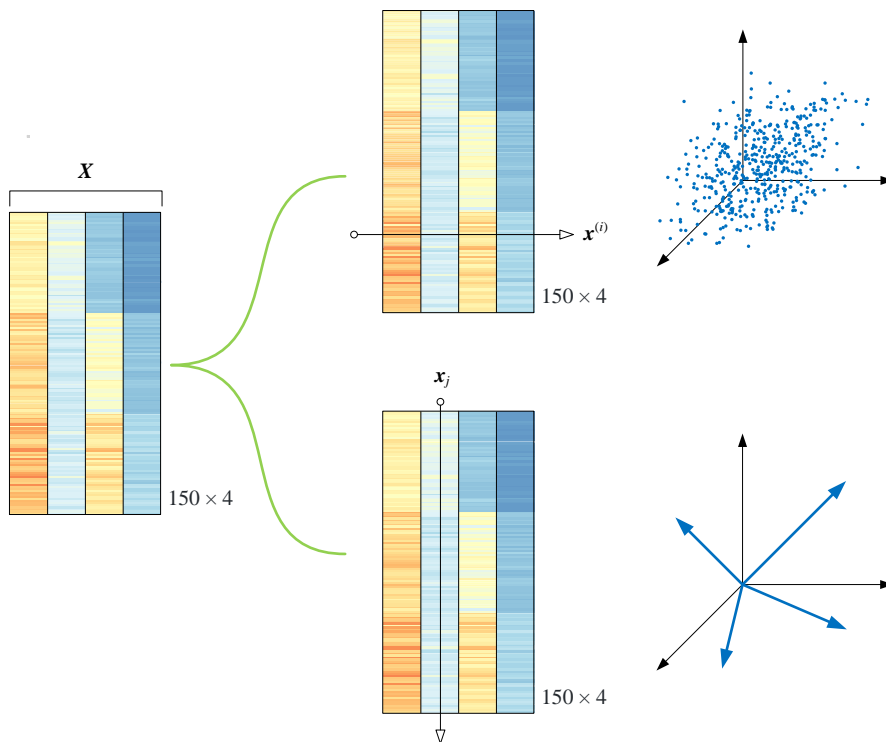


图 2. 矩阵可以分割成一系列行向量或列向量

## 图片

数据矩阵其实无处不在。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 为作者拍摄的一章鸢尾花照片。把这张照片做黑白处理后，它变成了形状为  $2990 \times 2714$  的矩阵，即 2990 行、2714 列。

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  $2 \times 2$  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。

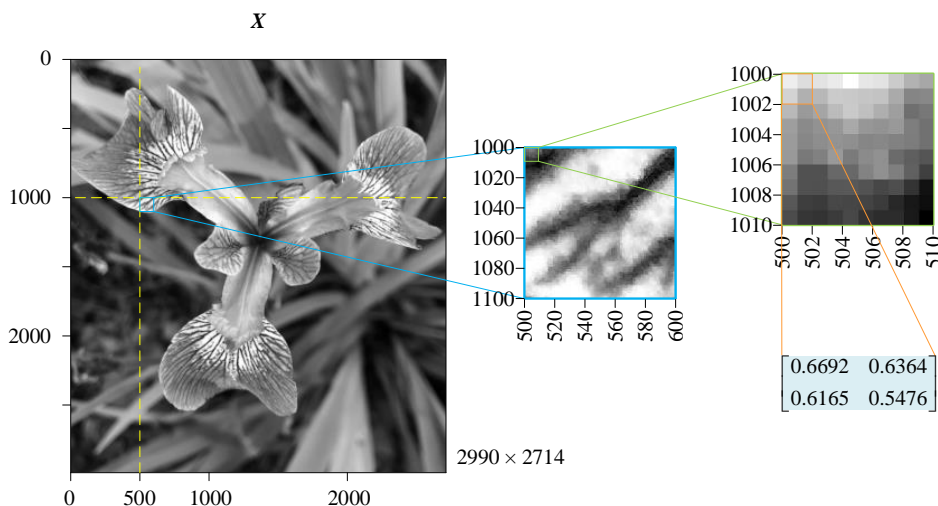


图 3. 照片也是数据矩阵

## 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， $i = 1 \sim 150$ 。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)。

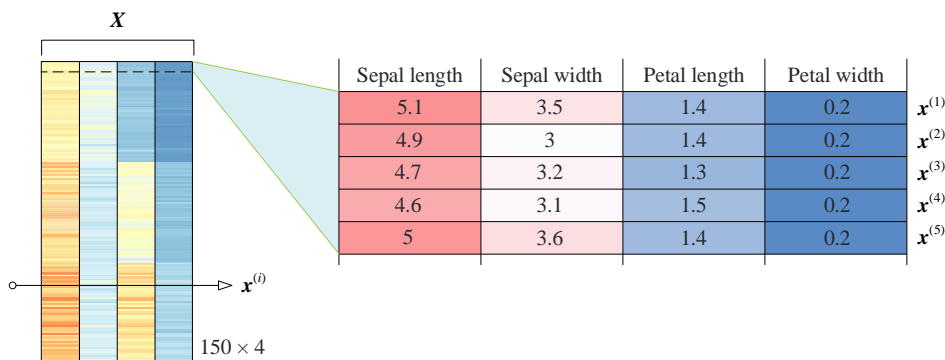


图 4. 鸢尾花数据，行向量代表样本数据点

比如， $\mathbf{X}$  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  $\mathbf{x}^{(1)}$  代表鸢尾花数据集中编号为 1 的样本。行向量  $\mathbf{x}^{(1)}$  可以视作 1 行、4 列的矩阵，即形状为  $1 \times 4$ 。

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)。长度、宽度度量单位为厘米 cm。

## 列向量

矩阵  $\mathbf{X}$  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  $\mathbf{X}$  的第  $j$  列可以写成行向量  $\mathbf{x}_j$ 。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， $\mathbf{X}$  的第 1 列行向量记做  $\mathbf{x}_1$ ，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  $150 \times 1$ 。

**⚠ 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  $\mathbf{x}_1$ 。

再次强调，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 有数据的地方，就有统计

前文提到，图 5 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  $\mathbf{x}_1$ ）、花萼宽度（第 2 列，列向量  $\mathbf{x}_2$ ）、花瓣长度（第 3 列，列向量  $\mathbf{x}_3$ ）和花瓣宽度（第 4 列，列向量  $\mathbf{x}_4$ ）。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。

从统计视角来看，我们可以计算样本数据各个特征的均值 ( $\mu_j$ )，可计算不同特征上样本数据的均方差 ( $\sigma_j$ )。有必要的话，我们还可以在图中给出  $\mu_j \pm \sigma_j$ 、 $\mu_j \pm 2\sigma_j$  对应的位置。图 5 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。

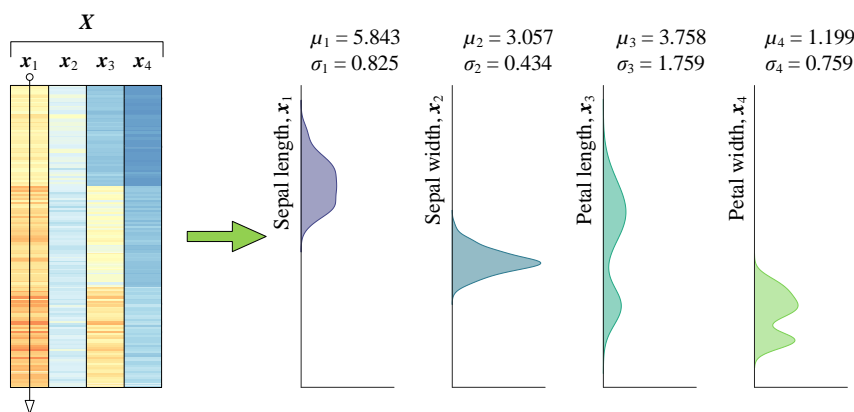


图 5. 鸢尾花数据，列向量代表数据特征

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

此外，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 6 所示为本书后续要介绍的鸢尾花数据矩阵衍生得到的几种矩阵。注意，图 2 和图 6 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 6 所示这些矩阵，本书第 24 章将探讨它们和各种矩阵分解 (matrix decomposition) 之间奇妙的关系。

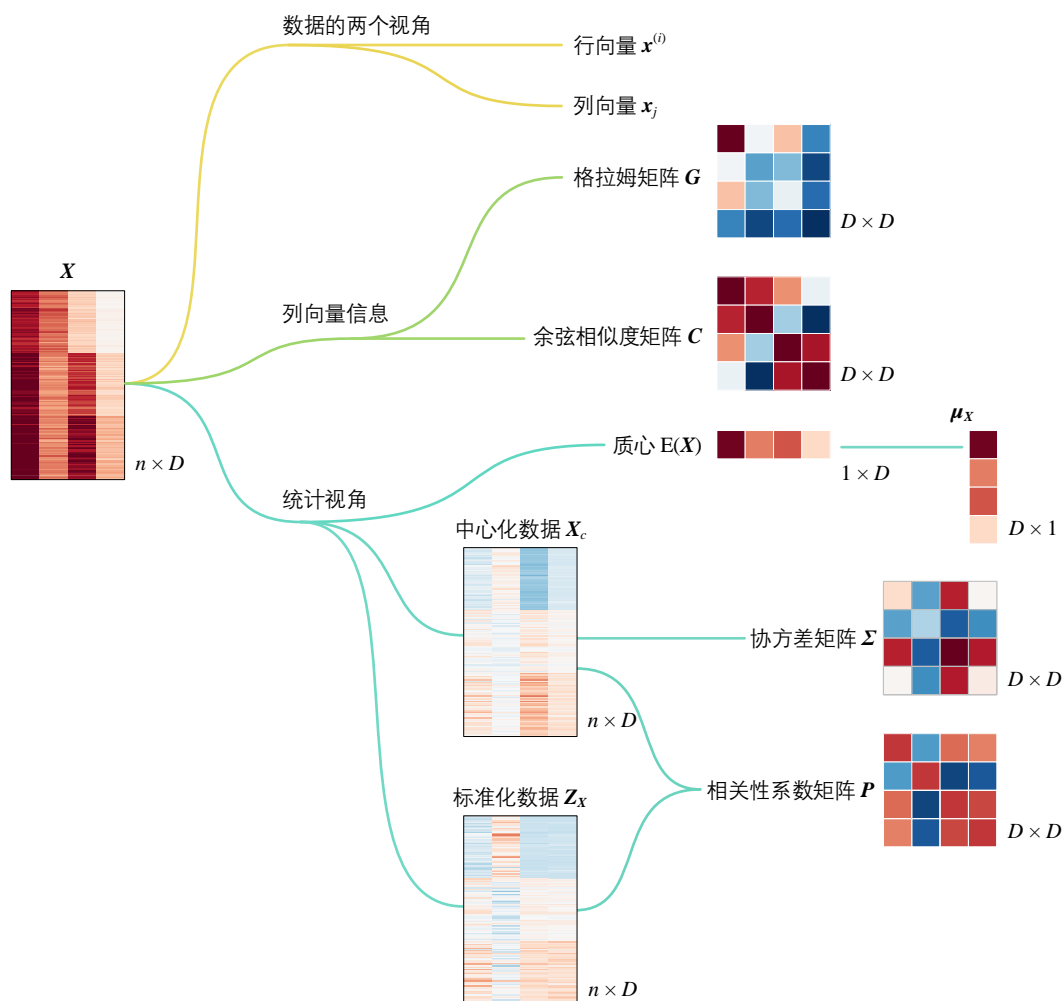


图 6. 鸢尾花数据衍生得到的几个矩阵

## 1.4 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 7 所示，这个图好比“样本数据云”。

图 7 中数据点 (5.0, 2.0) 可以写成行向量 [5.0, 2.0]。

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。



正交投影很好理解，即原来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，头顶正上方太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。

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

从集合视角来看，(5.0, 2.0) 属于  $\mathbb{R}^2$ ，即  $(5.0, 2.0) \in \mathbb{R}^2$ 。图 7 中整团数据云都属于  $\mathbb{R}^2$ 。

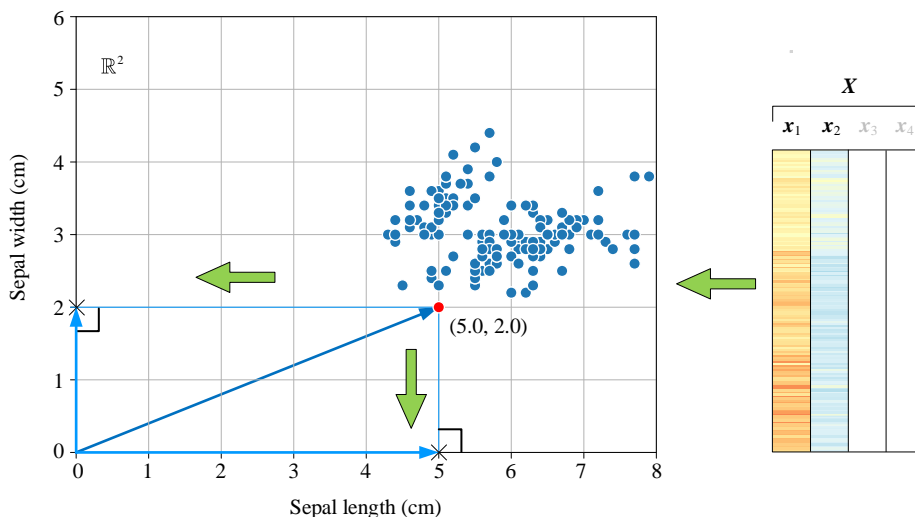


图 7. 鸢尾花前两个特征数据散点图

再者，如图 7 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，将整团数据云全部正交投影在横轴，得到图 8。图 8 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点。图 8 中横轴相当于一个一维空间，即一条数轴  $\mathbb{R}$ 。

我们也可以把整团数据云全部投影在纵轴，得到图 9。图中的  $\times$  是鸢尾花数据第二列花萼宽度数据。

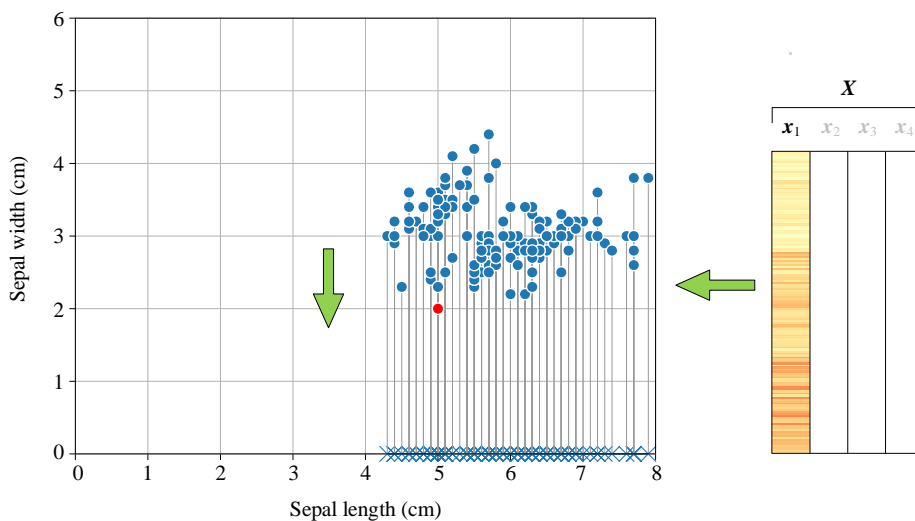


图 8. 二维散点正交投影到横轴

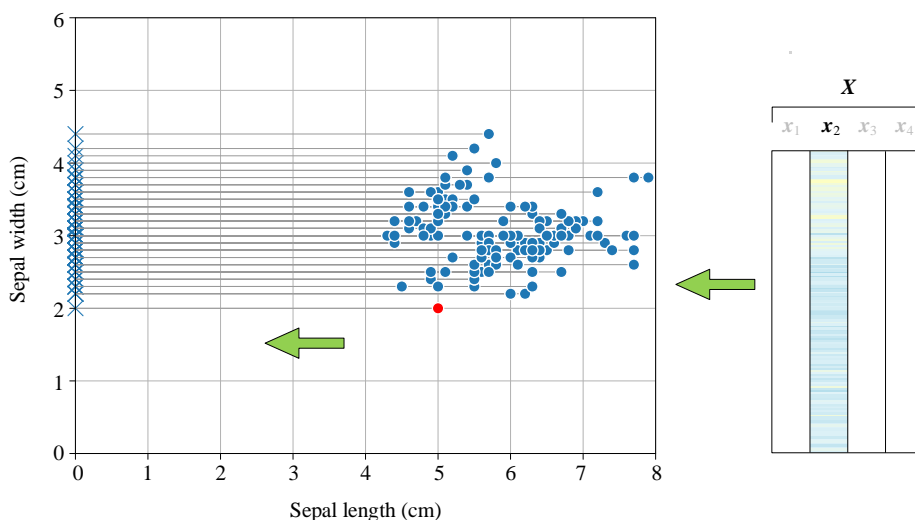


图 9. 二维散点正交投影到纵轴

你可能会问，是否可以将图 8 中所有点投影在一条斜线上？

答案是肯定的。

如图 10 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  $15^\circ$ 。观察图 10，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  $v_2x_2$ ，两种成分的组合  $v_1x_1 + v_2x_2$  就是投影点坐标。

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

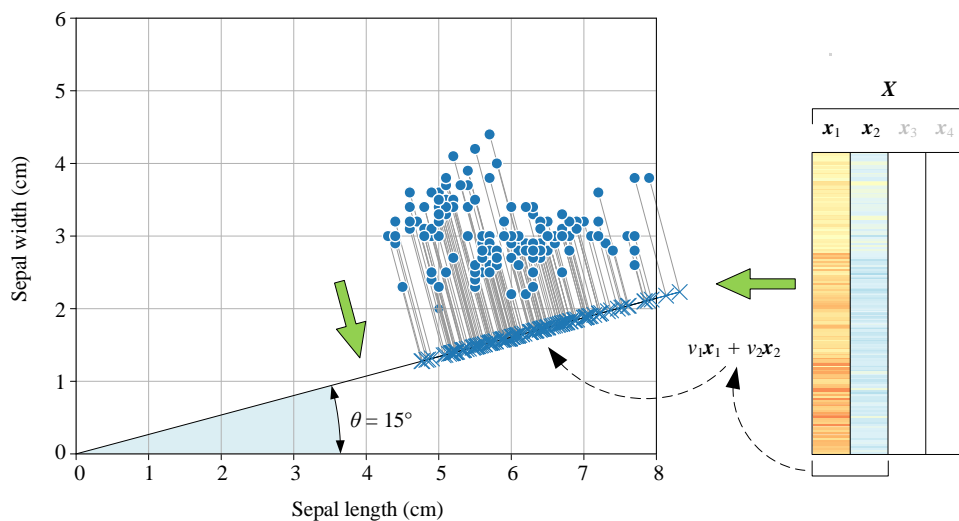


图 10. 二维散点正交投影到一条斜线

### 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  $\mathbb{R}^3$  绘制散点图, 得到图 11。这些原点  $\mathbf{0}$  和这些点的连线, 都代表行向量。图 7 相当于图 11 在水平面 (浅蓝色背景) 正交投影结果。

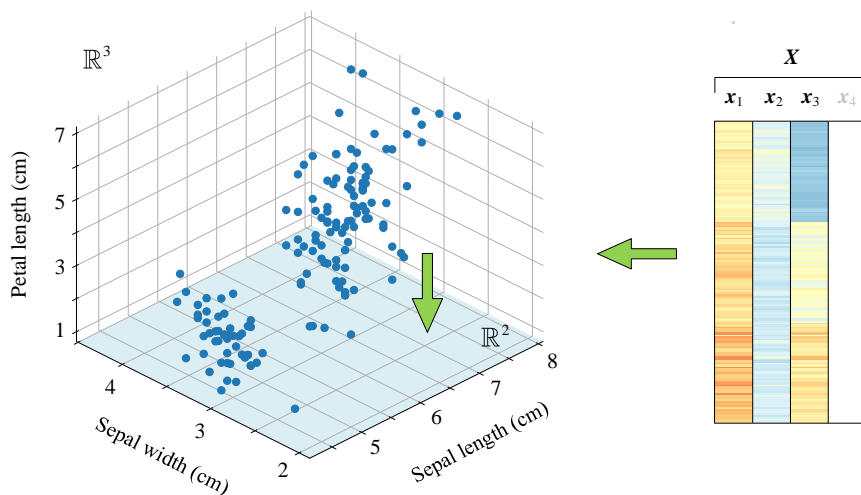


图 11. 鸢尾花前三个特征数据散点图

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 12 所示。成对特征散点图不但可以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 12 中的二维散点图相当于四维空间散点在不同平面上的投影结果。

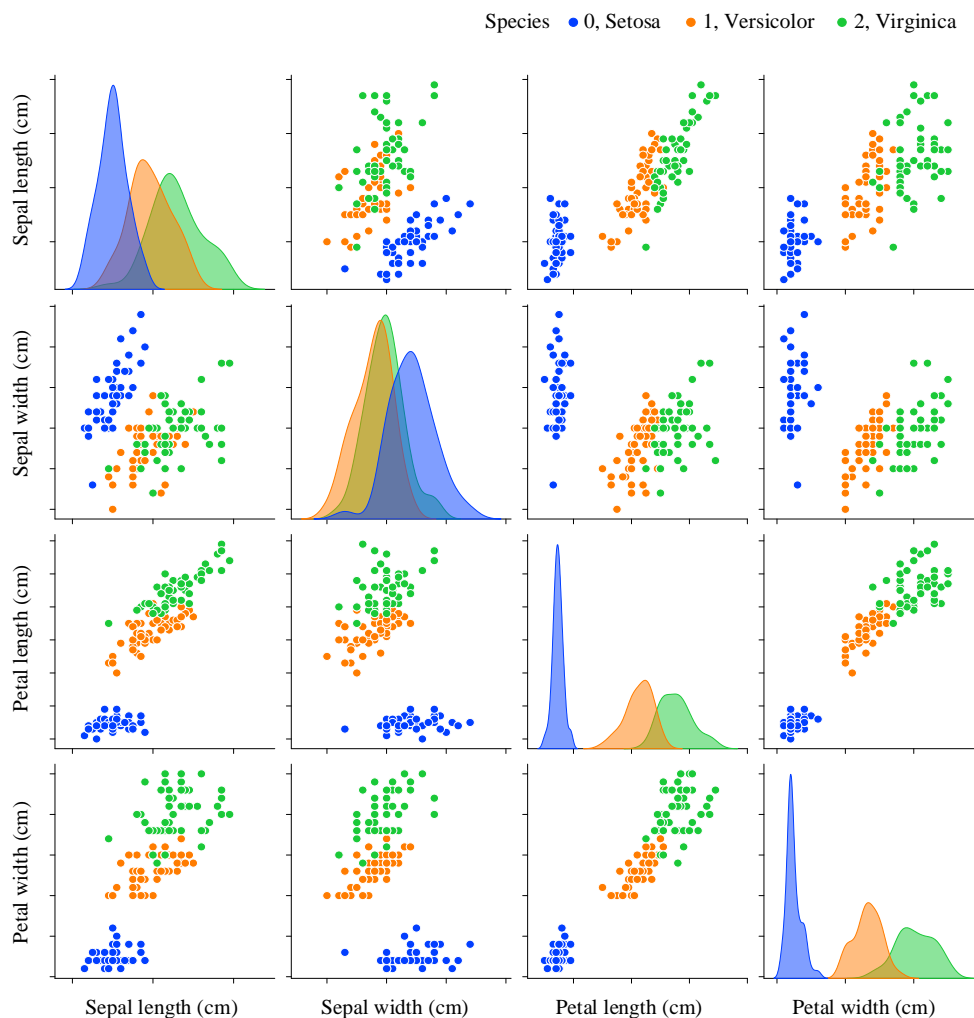


图 12. 鸢尾花数据成对特征散点图，考虑分类标签，图片来自《数学要素》

## 向量起点

如图 13 所示，本节前文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的**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 14 所示，将向量的起点移动到质心后，向量的长度、绝对角度、相对角度(向量两两之间的角度)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。

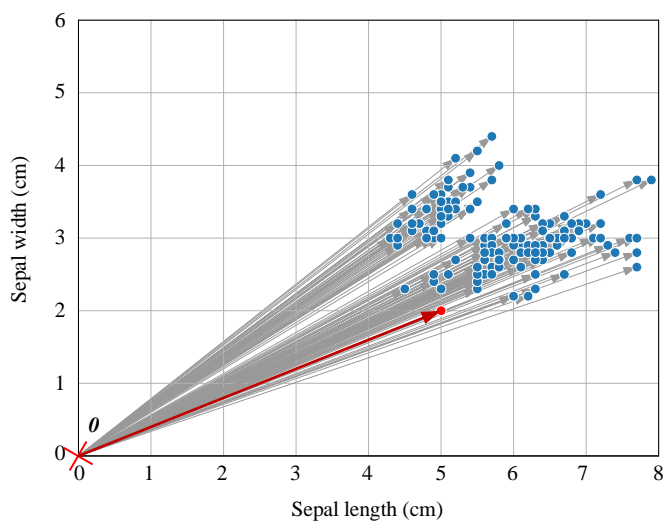


图 13. 向量起点为原点

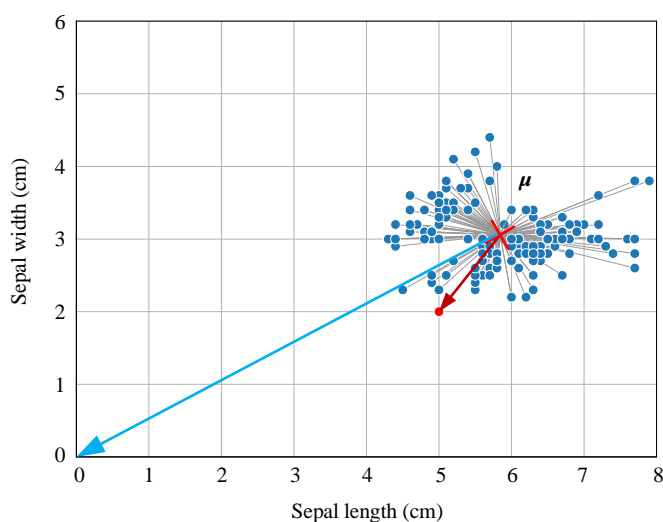



图 14. 向量起点为质心

将图 14 整团数据云随着质心平移到原点，这个过程就是去均值过程，结果如图 15 所示。去均值化得到的数据矩阵记做  $X_c$ ，显然  $X_c$  的质心位于原点  $\mathbf{0}$ 。

 观察图 12，我们发现，如果考虑数据标签的话，每一类都有自己质心，叫做分类质心，这是本书第 22 章要讨论的话题。

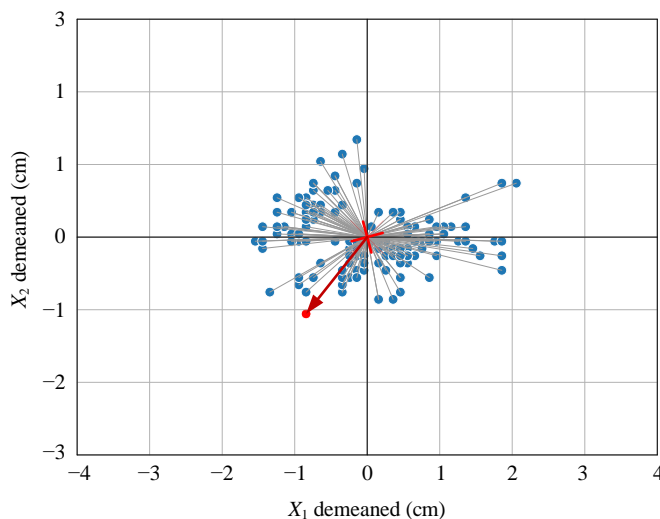


图 15. 数据去均值化

➡ 大家可能会问， $x_1$ 、 $x_2$ 、 $x_3$ 、 $x_4$ 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。

## 1.5 有向量的地方，肯定有空间

### 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)。我们在本系列丛书《数学要素》“鸡兔同笼三部曲”中，用线性方程组解决过鸡兔同笼问题。《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)：

$$\begin{cases} 1 \cdot x_1 + 1 \cdot x_2 = 35 \\ 2 \cdot x_1 + 4 \cdot x_2 = 94 \end{cases} \Rightarrow \underbrace{\begin{bmatrix} 1 & 1 \\ 2 & 4 \end{bmatrix}}_A \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_x = \underbrace{\begin{bmatrix} 35 \\ 94 \end{bmatrix}}_b \quad (3)$$

即：

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

未知变量构成的列向量  $x$  可以利用下式求得：

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



上述运算用到了矩阵乘法 (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 16 所示，矩阵  $A$  把  $e_1$  和  $e_2$  构成的方方正正的方格，变成一系列平行四边形网格，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  $A$  的两个列向量，即  $A = [a_1, a_2]$ 。

观察图 16 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 16 右图， $[a_1, a_2]$  同样张成了整个直角坐标系，不同的是网格形状。



本书将在第 7 章专门讲解向量空间。

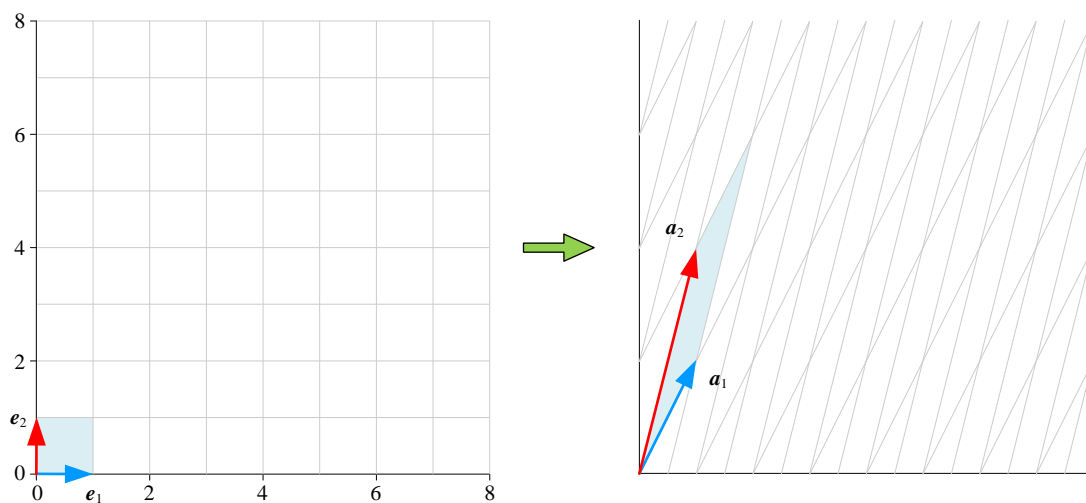


图 16. 矩阵  $A$  完成的线性变换

## 从正圆到旋转椭圆

椭圆等双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 17 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。单位圆上的向量  $x$ ，经过  $A$  的线性转换变成  $Ax$ 。

图 17 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  $-16.85^\circ$ 。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 18 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 章要探讨特征值分解。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。

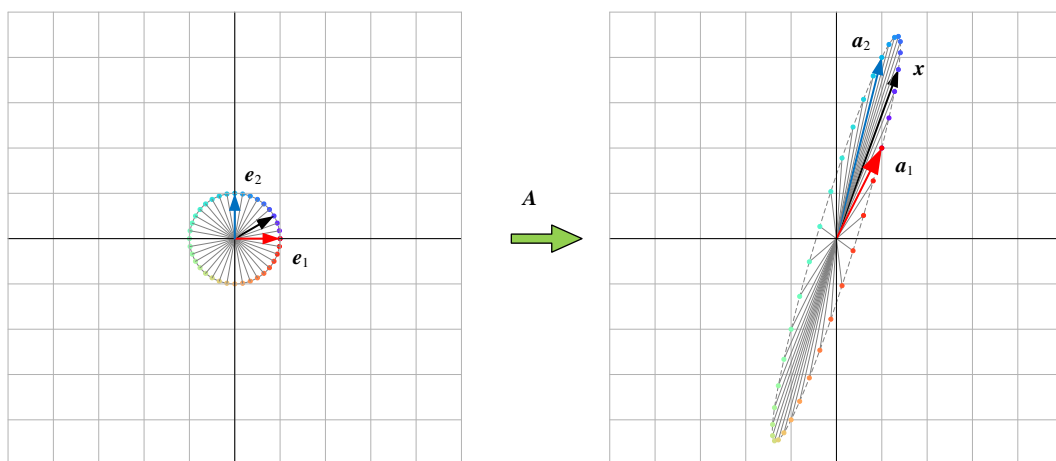


图 17. 矩阵  $A$  将单位圆转化为旋转椭圆

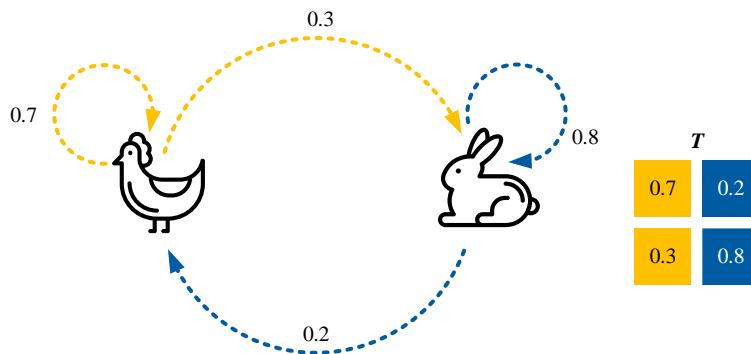
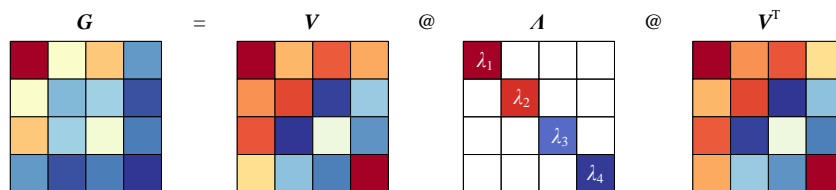


图 18. 鸡兔同笼三部曲中“鸡兔互变”，图片来自本系列丛书《数学要素》第 25 章

## 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 6 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 19 所示。请大家特别注意图 19 中的矩阵  $V$ 。正如图 16 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！

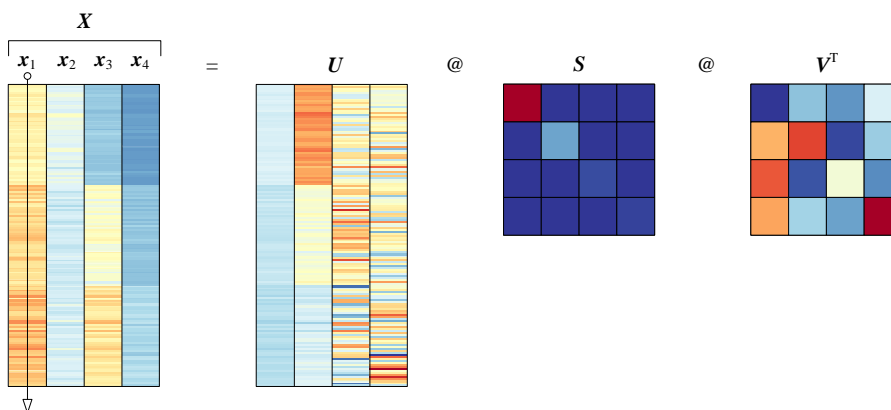


图 19. 矩阵  $X$  的格拉姆矩阵的特征值分解

## 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 20 所示为对鸢尾花数据矩阵的 SVD 分解，这幅图中的  $U$  和  $V$  都各自张起了不同的空间。

➔ 本书将在第 15、16 章专门讲解奇异值分解，并在第 23 章介绍 SVD 分解引出的四个空间。

图 20. 对矩阵  $X$  进行 SVD 分解

本章以向量为主线，蜻蜓点水地介绍了本书主要内容。不需要大家理解本章提到特所有术语，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！

有数据的地方，就有统计！

有向量的地方，就有几何！

有向量的地方，肯定有空间！