

# 22

## Statistics Meet Linear Algebra

# 数据与统计

有数据的地方，必有矩阵，亦必有统计



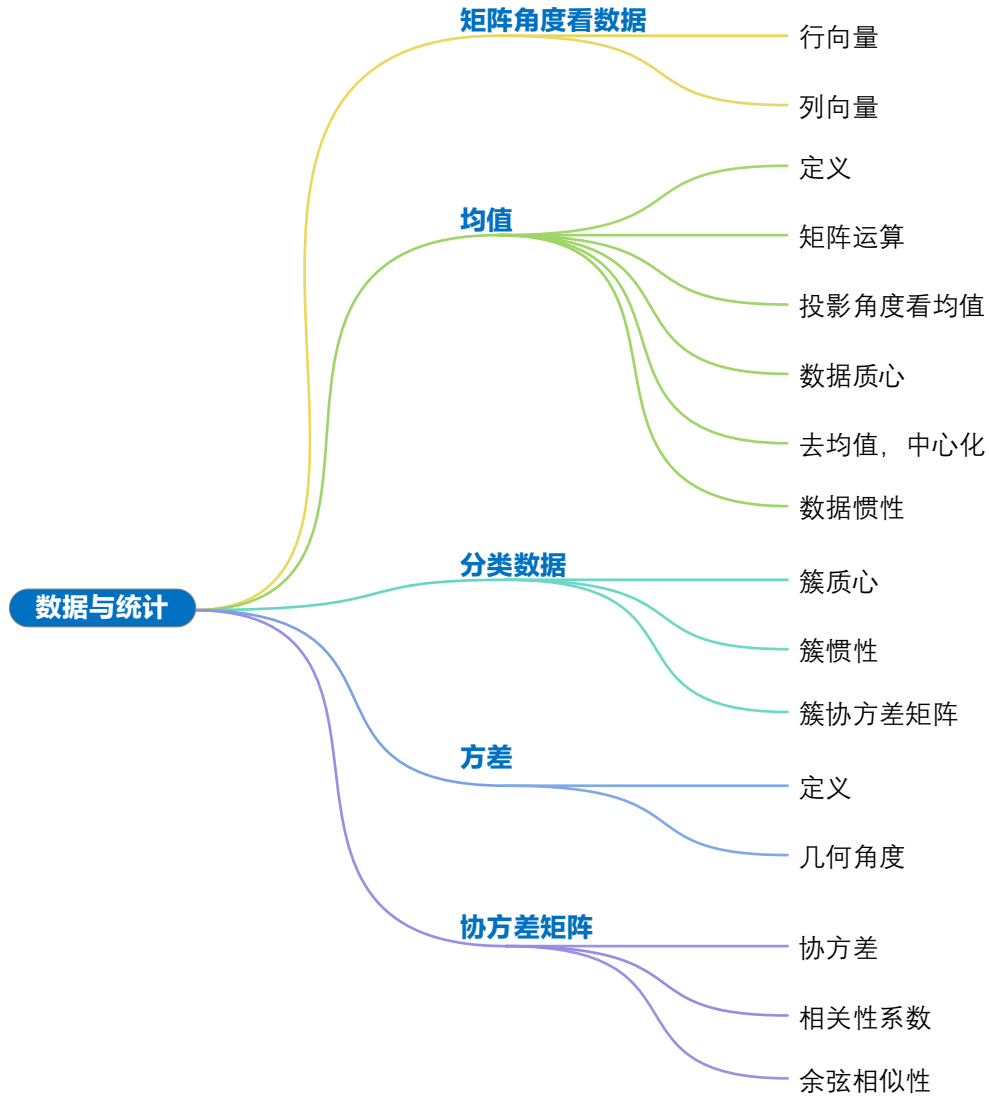
毫无争议的是，人类无法准确地判断事物的真伪，我们能做就是遵循更大的可能性。

***It is truth very certain that, when it is not in one's power to determine what is true, we ought to follow what is more probable.***

—— 勒内·笛卡尔 (René Descartes) | 法国哲学家、数学家、物理学家 | 1596 ~ 1650



- ◀ `numpy.linalg.norm()` 计算范数
- ◀ `numpy.ones()` 创建全 1 向量或全 1 矩阵
- ◀ `seaborn.heatmap()` 绘制热图
- ◀ `seaborn.kdeplot()` 绘制核密度估计曲线



## 22.1 统计 + 线性代数：以鸢尾花数据为例

本章大部分内容以鸢尾花数据为例，从线性代数运算视角讲解均值、方差、协方差、相关性系数、协方差矩阵、相关性系数矩阵等统计相关知识点。

### 鸢尾花数据集

回顾鸢尾花数据集，不考虑鸢尾花品种，数据矩阵  $X$  的形状为  $150 \times 4$ ，即 150 行、4 列。

鸢尾花数据集共有四个特征——花萼长度、花萼宽度、花瓣长度和花瓣宽度。这些特征依次对应  $X$  的四列。图 1 所示为用热图可视化鸢尾花数据集。数据的每一行代表一朵花，每一列代表一个特征上的所有数据。

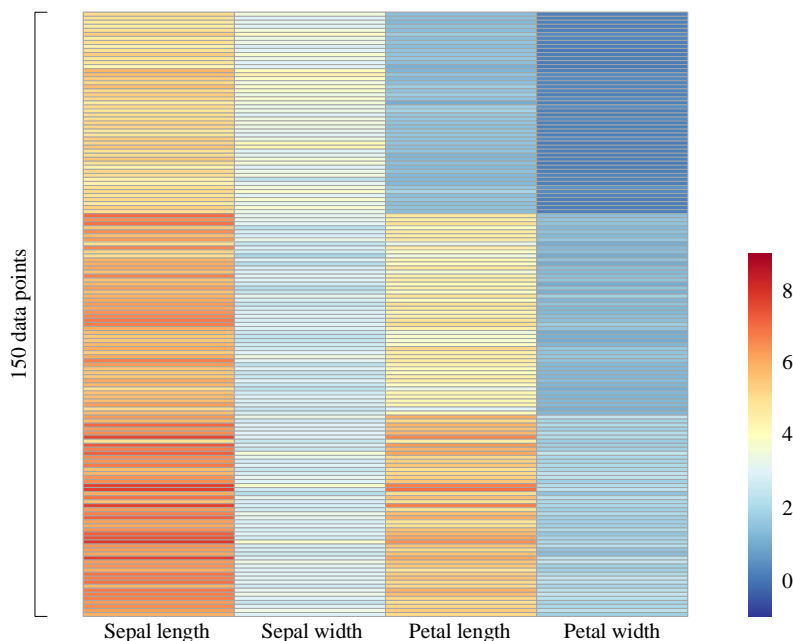


图 1. 鸢尾花数据，原始数据矩阵  $X$ ，单位为厘米 (cm)



Bk4\_Ch22\_01.py 中 Bk4\_Ch22\_01\_A 部分绘制图 1。

## 22.2 均值：线性代数视角

从样本数据矩阵  $\mathbf{X}$  中，取出任意一列列向量  $\mathbf{x}_j$ 。 $\mathbf{x}_j$  代表着第  $j$  特征的所有样本数据构成的列向量：

$$\mathbf{x}_j = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix} \quad (1)$$

列向量  $\mathbf{x}_j$  对应随机变量  $X_j$ 。

通过样本数据估算随机变量  $X_j$  的期望值 (均值)  $E(X_j)$ ：

$$E(X_j) = \mu_j = \frac{x_{1,j} + x_{2,j} + \cdots + x_{n,j}}{n} = \frac{1}{n} \sum_{i=1}^n x_{i,j} \quad (2)$$

▲ 注意，(2) 上式中  $1/n$  为权重。计算均值时，(2) 中每个数据点为等概率。我们以后还会遇到加权平均值 (weighted average)，也就是说计算均值时不同的数据点权重不同。

本书中， $E(X_j)$  等价于  $E(\mathbf{x}_j)$ 。 $E(\mathbf{x}_j)$  对应的线性代数运算如下：

$$E(\mathbf{x}_j) = E(X_j) = \mu_j = \frac{\mathbf{x}_j^T \mathbf{1}}{n} = \frac{\mathbf{1}^T \mathbf{x}_j}{n} = \frac{\mathbf{x}_j \cdot \mathbf{1}}{n} = \frac{\mathbf{1} \cdot \mathbf{x}_j}{n} = \frac{1}{n} \sum_{i=1}^n x_{i,j} \quad (3)$$

其中， $\mathbf{1}$  为全 1 列向量，行数和  $\mathbf{x}_j$  一致。

(3) 左乘  $n$  可以得到如下等式：

$$n\mu_j = nE(\mathbf{x}_j) = \mathbf{x}_j^T \mathbf{1} = \mathbf{1}^T \mathbf{x}_j = \mathbf{x}_j \cdot \mathbf{1} = \mathbf{1} \cdot \mathbf{x}_j \quad (4)$$

图 2 所示为计算  $E(\mathbf{x}_j)$  对应的矩阵运算示意图。

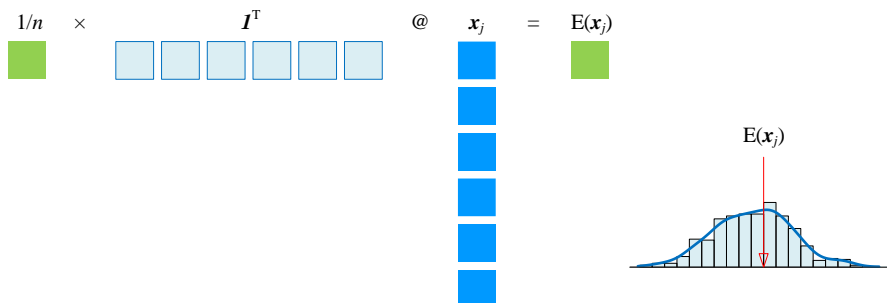


图 2. 计算  $\mathbf{x}_j$  期望值/均值

利用矩阵运算分别得到鸢尾花的四个特征的期望值：

$$\begin{cases} E(\mathbf{x}_1) = \mu_1 = 5.843 \\ \text{Sepal length} \\ E(\mathbf{x}_2) = \mu_2 = 3.057 \\ \text{Sepal width} \\ E(\mathbf{x}_3) = \mu_3 = 3.758 \\ \text{Petal length} \\ E(\mathbf{x}_4) = \mu_4 = 1.199 \\ \text{Petal width} \end{cases} \quad (5)$$

## 向量视角

下面我们聊一聊解释  $E(\mathbf{x}_j)$  的有趣角度——投影。

如图 3 所示， $E(\mathbf{x}_j)$  是一个标量，而向量  $E(\mathbf{x}_j)\mathbf{I}$  相当于向量  $\mathbf{x}_j$  在  $\mathbf{I}$  方向上投影的向量投影结果：

$$E(\mathbf{x}_j)\mathbf{I} = \text{proj}_{\mathbf{I}}(\mathbf{x}_j) = \frac{\mathbf{x}_j^T \mathbf{I}}{\mathbf{I}^T \mathbf{I}} \mathbf{I} = \frac{\mathbf{x}_j^T \mathbf{I}}{n} \mathbf{I} \quad (6)$$

再次注意， $E(\mathbf{x}_j)$  为标量； $E(\mathbf{x}_j)\mathbf{I}$  为向量，和  $\mathbf{I}$  平行。

图 3 中， $\mathbf{I}$  方向上解释了  $\mathbf{x}_j$  中  $E(\mathbf{x}_j)\mathbf{I}$  这部分分量，没有被解释的向量分量为：

$$\mathbf{x}_j - \text{proj}_{\mathbf{I}}(\mathbf{x}_j) = \mathbf{x}_j - E(\mathbf{x}_j)\mathbf{I} \quad (7)$$

(7) 这部分垂直于  $\mathbf{I}$ ，也就是说：

$$\mathbf{I}^T (\mathbf{x}_j - \text{proj}_{\mathbf{I}}(\mathbf{x}_j)) = \mathbf{I}^T \left( \mathbf{x}_j - \frac{\mathbf{x}_j^T \mathbf{I}}{n} \mathbf{I} \right) = \mathbf{I}^T \mathbf{x}_j - \frac{\mathbf{x}_j^T \mathbf{I}}{n} \mathbf{I}^T \mathbf{I} = \mathbf{I}^T \mathbf{x}_j - \mathbf{x}_j^T \mathbf{I} = 0 \quad (8)$$

注意，上式中  $\mathbf{x}_j^T \mathbf{I}$  为标量，因此  $\mathbf{I}^T (\mathbf{x}_j^T \mathbf{I}) \mathbf{I} = (\mathbf{x}_j^T \mathbf{I}) \mathbf{I}^T \mathbf{I}$ 。均值作为一个统计量，它能解释列向量  $\mathbf{x}_j$  一部分特征。 $\mathbf{x}_j - E(\mathbf{x}_j)\mathbf{I}$  将在标准差（方差平方根）中加以解释。

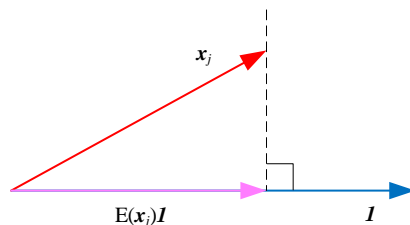


图 3. 投影角度看期望值

## 两个极端例子

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](mailto:jiang.visualize.ml@gmail.com)

如果  $\mathbf{x}_j$  所有元素均相同，比如全都是  $k$ ，那么  $\mathbf{x}_j$  可以写成：

$$\mathbf{x}_j = \begin{bmatrix} k \\ k \\ \vdots \\ k \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = k\mathbf{I} \quad (9)$$

这种情况， $\mathbf{x}_j$  和  $\mathbf{I}$  共线。

再举个相反的例子，如果  $\mathbf{x}_j$  和  $\mathbf{I}$  垂直，

$$\mathbf{I}^\top \mathbf{x}_j = 0 \quad (10)$$

也就是意味着  $E(\mathbf{x}_j) = 0$ 。也就是说， $\mathbf{x}_j$  在  $\mathbf{I}$  方向的标量投影为 0。

➔ 对于最小二乘法线性回归， $\mathbf{x}_j - E(\mathbf{x}_j)\mathbf{I}$  垂直于  $\mathbf{I}$  这一结论格外重要。本系列丛书《数据科学》将深入讨论如何用向量视角解释最小二乘法线性回归。

## 22.3 质心：均值排列成向量

上一节，我们探讨了一个特征的均值，本节介绍数据矩阵  $\mathbf{X}$  的每列特征均值构成的向量，我们管这个向量叫做数据的**质心** (centroid)。图 4 所示为平面上数据  $\mathbf{X}$  的质心位置。

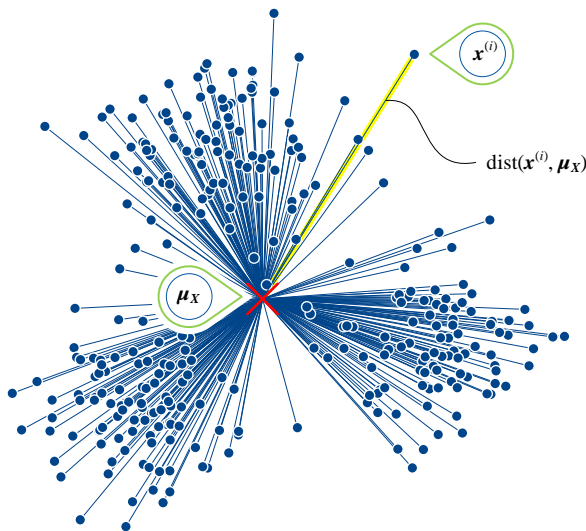


图 4. 平面上数据矩阵  $\mathbf{X}$  质心位置

### 列向量

$X$  样本数据的质心  $\mu_X$  定义如下：

$$\mu_X = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} = \begin{bmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_D) \end{bmatrix} \quad (11)$$

⚠ 注意，为了方便运算， $\mu_X$  被定义为列向量。

比如，在多元高斯分布中，我们会用到列向量  $\mu_X$ 。比如，多元高斯分布的概率密度函数：

$$f_X(x) = \frac{\exp\left(-\frac{1}{2}(x - \mu_X)^T \Sigma^{-1}(x - \mu_X)\right)}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \quad (12)$$

上式中，几何角度来看， $x - \mu_X$  相当于“平移”， $\Sigma^{-1}$  则提供“缩放 + 旋转”。对这部分内容感到生疏的读者，请回顾本书第 20 章。

前文介绍过， $\mu_X$  可以通过如下矩阵运算获得：

$$\mu_X = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} = \frac{(I^T X)^T}{n} = \frac{X^T I}{n} \quad (13)$$

其中，样本数据矩阵  $X$  为  $n$  行、 $D$  列矩阵，即有  $n$  个样本， $D$  个特征。

整理 (13) 得到两个等式：

$$\begin{cases} X^T I = n\mu_X \\ I^T X = n(\mu_X)^T \end{cases} \quad (14)$$

举个例子，鸢尾花数据质心位置：

$$\mu_X = \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \quad (15)$$

本书第 5 章讲过上述内容。

## 行向量

为了区分，丛书特别定义  $E(X)$  为行向量，即：

$$\begin{aligned}
 E(\mathbf{X}) &= [E(\mathbf{x}_1) \ E(\mathbf{x}_2) \ \cdots \ E(\mathbf{x}_D)] \\
 &= [\mu_1 \ \mu_2 \ \cdots \ \mu_D] \\
 &= (\boldsymbol{\mu}_X)^T = \frac{\mathbf{I}^T \mathbf{X}}{n}
 \end{aligned} \tag{16}$$

整理 (16)，可以得到：

$$\mathbf{I}^T \mathbf{X} = nE(\mathbf{X}) \tag{17}$$

图 5 所示为计算质心示意图，以及  $E(\mathbf{X})$  和  $\boldsymbol{\mu}_X$  之间关系。

$E(\mathbf{X})$  一般用在和数据矩阵  $\mathbf{X}$  相关的计算中，比如中心化 (去均值)  $\mathbf{X} - E(\mathbf{X})$ 。 $\mathbf{X} - E(\mathbf{X})$  用到了本书第 4 章介绍的“广播原则”。

⚠ 注意，本系列丛书中， $E(\boldsymbol{\chi})$  仍然为列向量。 $\boldsymbol{\chi}$  代表  $X_1, X_2 \dots$  等随机变量构成的列向量。 $E(\bullet)$  为求期望值运算符，作用于列向量  $\boldsymbol{\chi}$ ，结果还是列向量。而  $\mathbf{X}$  的每一列代表一个随机变量， $E(\bullet)$  作用于数据矩阵  $\mathbf{X}$  时， $E(\mathbf{X})$  为行向量。

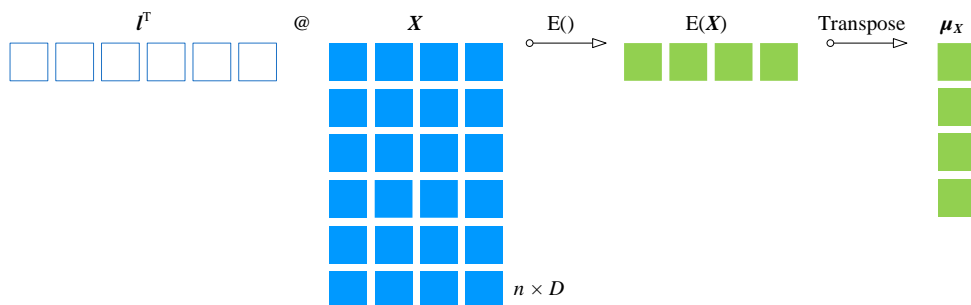


图 5. 计算  $\mathbf{X}$  样本数据的质心  $\boldsymbol{\mu}_X$

## 22.4 中心化：平移

### 中心化、去均值

数据矩阵  $\mathbf{X}$  中第  $j$  特征特征数据  $\mathbf{x}_j$  减去其均值  $\mu_j$ ，对应的矩阵运算为：

$$\mathbf{x}_j - \mathbf{I}\mu_j = \mathbf{x}_j - \underbrace{\frac{1}{n}\mathbf{I}\mathbf{I}^T}_{\mathbf{M}}\mathbf{x}_j = \left(\mathbf{I} - \frac{1}{n}\mathbf{I}\mathbf{I}^T\right)\mathbf{x}_j \tag{18}$$

上式没有使用“广播原则”。其中， $\mathbf{I}\mathbf{I}^T$  为全 1 列向量和其转置乘积，结果为方阵。

而数据矩阵  $\mathbf{X}$  中每一列数据  $\mathbf{x}_j$  分别减去对应本列均值  $\mu_j$  得到  $\mathbf{X}_c$ ，对应矩阵运算为：



$$\mathbf{X}_c = \mathbf{X} - \mathbf{I} \left( \frac{\mathbf{X}^T \mathbf{I}}{n} \right)^T = \mathbf{X} - \frac{1}{n} \mathbf{I} \mathbf{I}^T \mathbf{X} = \underbrace{\left( \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T \right)}_{\mathbf{M}} \mathbf{X} \quad (19)$$

我们管这个运算叫做数据**中心化** (centralize)，也叫**去均值** (demean)。

令：

$$\mathbf{M} = \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T \quad (20)$$

本章后文称  $\mathbf{M}$  为中心化矩阵，或去均值矩阵。

为了方便，我们一般利用广播原则来中心化  $\mathbf{X}$ ，即  $\mathbf{X}$  减去行向量  $\mathbf{E}(\mathbf{X})$  得到  $\mathbf{X}_c$ ：

$$\mathbf{X}_c = \mathbf{X} - \mathbf{E}(\mathbf{X}) \quad (21)$$

中心化后，数据  $\mathbf{X}_c$  质心位于原点  $\mathbf{0}$ 。

## 中心化矩阵

我们在 (18) 和 (19) 都看到了中心化矩阵  $\mathbf{M}$ ，下面我们简单分析一下这个特殊矩阵。

将  $\mathbf{M}$  展开得到：

$$\mathbf{M} = \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} - \begin{bmatrix} 1/n & 1/n & \cdots & 1/n \\ 1/n & 1/n & \cdots & 1/n \\ \vdots & \vdots & \ddots & \vdots \\ 1/n & 1/n & \cdots & 1/n \end{bmatrix} = \begin{bmatrix} 1-1/n & -1/n & \cdots & -1/n \\ -1/n & 1-1/n & \cdots & -1/n \\ \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & \cdots & 1-1/n \end{bmatrix} \quad (22)$$

矩阵  $\mathbf{M}$  为对称矩阵， $\mathbf{M}$  的主对角线元素为  $1 - 1/n$ ，剩余元素为  $-1/n$ 。

矩阵  $\mathbf{M}$  为幂等矩阵，即满足：

$$\mathbf{M} \mathbf{M} = \mathbf{M} \quad (23)$$

将 (20) 代入上式，展开整理：

$$\begin{aligned} \mathbf{M} \mathbf{M} &= \left( \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T \right) \left( \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T \right) = \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T + \frac{1}{n} \mathbf{I} \mathbf{I}^T \frac{1}{n} \mathbf{I} \mathbf{I}^T \\ &= \mathbf{I} - \frac{2}{n} \mathbf{I} \mathbf{I}^T + \frac{1}{n} \mathbf{I} \mathbf{I}^T = \mathbf{I} - \frac{1}{n} \mathbf{I} \mathbf{I}^T = \mathbf{M} \end{aligned} \quad (24)$$

我们在后文还会用到  $\mathbf{M}$  这个中心化矩阵。

(24) 中所有全 1 列向量  $\mathbf{I}$  等长，形状均为  $n \times 1$ 。因此  $\mathbf{I} \mathbf{I}^T$  结果为  $n \times n$  方阵，矩阵中每个元素都是 1。而  $\mathbf{I}^T \mathbf{I}$  结果为标量  $n$ 。我们也会在很多运算中看到  $\mathbf{I} \mathbf{I}^T$  中两个  $\mathbf{I}$  长度不同。此时， $\mathbf{I} \mathbf{I}^T$  结果为长方阵。此外，(24) 中两个单位矩阵  $\mathbf{I}$  也都是  $n \times n$  方阵。大家遇到单位矩阵时要注意其形状，比如  $\mathbf{I} \mathbf{A}_{m \times n} \mathbf{I} = \mathbf{A}_{m \times n}$  这个等式左右的单位矩阵形状显然不同，左边  $\mathbf{I}$  形状为  $m \times m$ ，右边  $\mathbf{I}$  形状为  $n \times n$ 。

### 标准化：平移 + 缩放

在中心化的基础上，我们可以进一步对  $X_c$  进行**标准化** (standardization 或 z-score normalization)。计算过程为，对原始数据先去均值，然后每一列再除以对应标准差。对应的矩阵运算如下：

$$\mathbf{Z}_X = \mathbf{X}_c \mathbf{S}^{-1} = (\mathbf{X} - \mathbf{E}(\mathbf{X})) \mathbf{S}^{-1} \quad (25)$$

其中，缩放矩阵  $\mathbf{S}$  为：

$$\mathbf{S} = \text{diag}(\text{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix} \quad (26)$$

其中，里层  $\text{diag}()$  提取对角线元素，结果为向量；外层  $\text{diag}()$  将向量展成对角方阵。

(25) 处理得到的数值实际上是原始数据的 **z 分数** (z score)，含义是距离均值若干倍的标准差偏移。比如说，标准化得到的数值为 3，也就是说这个数据距离均值 3 倍标准差偏移。数值的正负表达偏移的方向。

**▲ 注意**，数据标准化过程也是一个“去单位化”过程。去单位数值有利于联系、比较单位不同、取值范围差异较大的样本数据。此外，本章不会区分总体标准差和样本标准差记号。

### 惯性

数据**惯性** (inertia) 可以用来描述样本数据紧密程度，惯性实际上就是**总离差平方和** (Sum of Squares for Deviations, SSD)，定义如下：

$$\text{SSD}(\mathbf{X}) = \sum_{i=1}^n \text{dist}(\mathbf{x}^{(i)}, \mathbf{E}(\mathbf{X}))^2 = \sum_{i=1}^n \|\mathbf{x}^{(i)} - \mathbf{E}(\mathbf{X})\|_2^2 = \sum_{i=1}^n \|\mathbf{x}^{(i)\top} - \boldsymbol{\mu}_X\|_2^2 \quad (27)$$

如图 4 所示，SSD 相当于样本点和质心  $\mathbf{E}(\mathbf{X})$  欧氏距离平方和。

(27) 相当于中心化数据  $\mathbf{X}_c$  每个行向量和自身求内积后，再求和。用迹  $\text{trace}()$  可以方便得到 SSD 结果：

$$\text{SSD}(\mathbf{X}) = \text{trace}(\mathbf{X}_c^\top \mathbf{X}_c) = \text{trace}((\mathbf{X} - \mathbf{E}(\mathbf{X}))^\top (\mathbf{X} - \mathbf{E}(\mathbf{X}))) \quad (28)$$



Bk4\_Ch22\_01.py 中 Bk4\_Ch22\_01\_B 部分绘制图 6 并计算 SSD。请大家根据本节代码自行计算并绘制标准化鸢尾花数据热图。

## 22.5 分类数据：加标签

大家都清楚鸢尾花样本数据有三类标签，定义为  $C_1$ 、 $C_2$ 、 $C_3$ ，具体如图 7 所示。

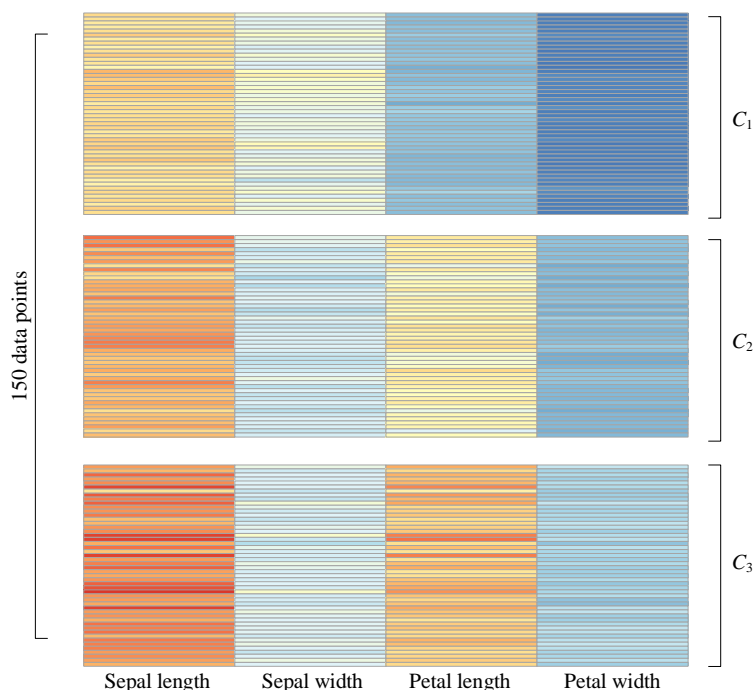


图 7. 鸢尾花数据分为三类

### 簇质心

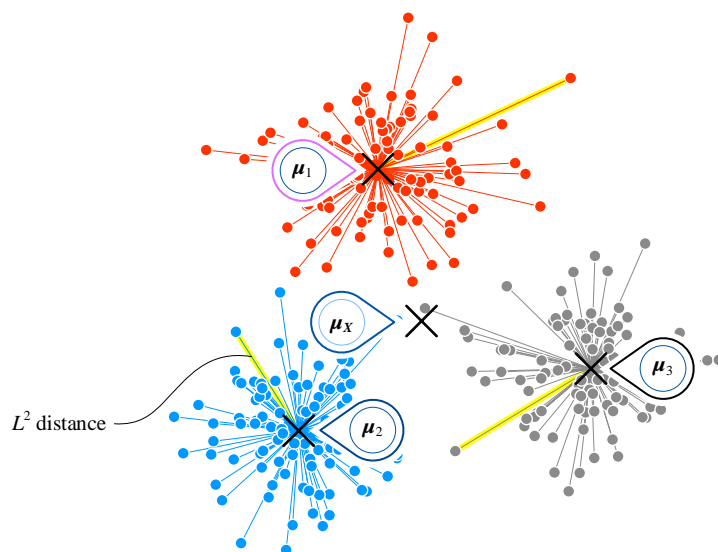
类似  $\mu_X$ ，任意一类标签为  $C_k$  样本数据的簇质心  $\mu_k$ ，定义如下：

$$\mu_k = \frac{1}{\text{count}(C_k)} \sum_{i \in C_k} \mathbf{x}^{(i)\top} \quad (29)$$

公式看上去复杂，道理其实很简单。

翻译一下，对于属于某个标签  $C_k$  的所有样本数据  $\mathbf{x}^{(i)}$  ( $i \in C_k$ )，求其各个特征平均值，构造成一个新的列向量  $\mu_k$ 。图 8 所示为样本数据质心  $\mu_X$ ，和三个不同标签数据各自的簇质心  $\mu_1$ 、 $\mu_2$  和  $\mu_3$  之间的关系。

⚠ 注意， $\mathbf{x}^{(i)}$  为行向量，而  $\mu_k$  为列向量。这就是为什么 (29) 存在转置运算。

图 8. 样本数据质心  $\mu_X$ ，和三类数据各自的质心  $\mu_1$ 、 $\mu_2$  和  $\mu_3$ 

### 举个例子

假设样本数据中只有第 2、5、6 和 9 四个数据点标签为  $C_1$ ，它们构成了原始数据的一个子集： $\{(\mathbf{x}^{(2)}, y^{(2)} = C_1), (\mathbf{x}^{(5)}, y^{(5)} = C_1), (\mathbf{x}^{(6)}, y^{(6)} = C_1), (\mathbf{x}^{(9)}, y^{(9)} = C_1)\}$ 。

数据点有两特征，具体坐标值如下：

$$\mathbf{x}^{(2)} = [2 \ 3], \mathbf{x}^{(5)} = [3 \ 1], \mathbf{x}^{(6)} = [-2 \ 2], \mathbf{x}^{(9)} = [1 \ 6] \quad (30)$$

则标签为  $C_1$  簇质心位置为  $[1, 3]^T$ ，具体运算过程如下：

$$\begin{aligned} \mu_{C_1} &= \frac{1}{\text{count}(C_1)} \sum_{i \in C_1} \mathbf{x}^{(i)T} = \frac{1}{\text{count}(C_1)} (\mathbf{x}^{(2)T} + \mathbf{x}^{(5)T} + \mathbf{x}^{(6)T} + \mathbf{x}^{(9)T}) \\ &= \frac{1}{4} ([2 \ 3]^T + [3 \ 1]^T + [-2 \ 2]^T + [1 \ 6]^T) = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \end{aligned} \quad (31)$$

以鸢尾花数据为例，计算簇质心就是对图 7 三组标签不同样本数据分别计算质心。图 9 不同颜色的  $\times$  代表不同标签鸢尾花的簇质心位置。

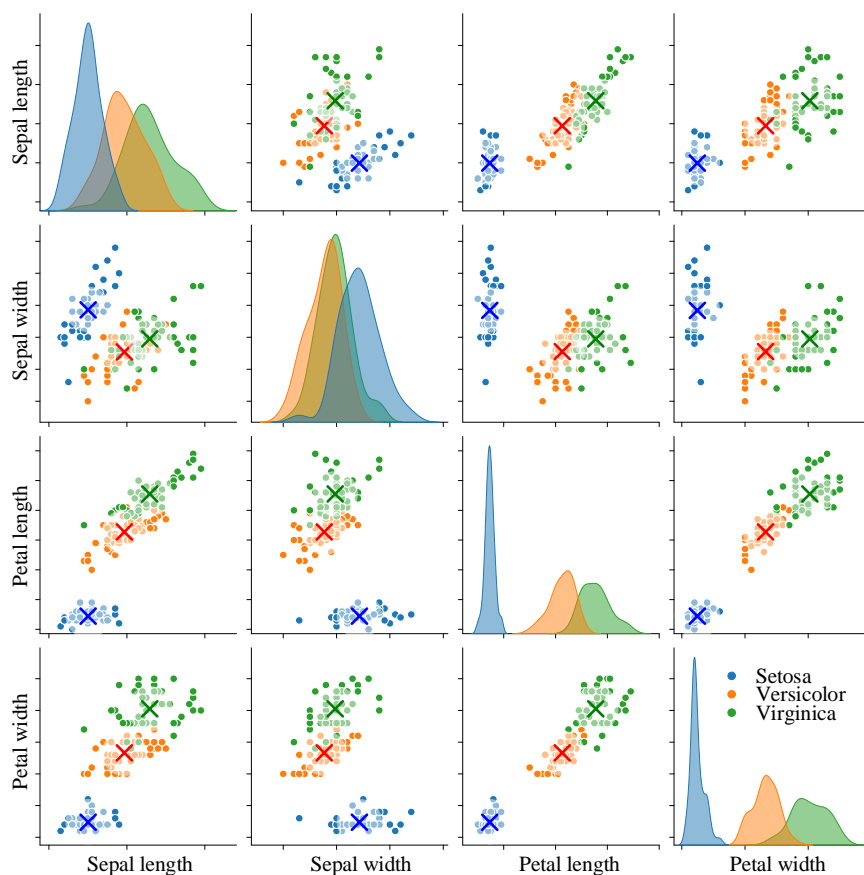


图 9. 鸢尾花数据簇质心位置

## 22.6 方差：均值向量没有解释的部分

对于总体来说，随机变量  $X$  方差的计算式为：

$$\text{var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - E(X))^2 \quad (32)$$

注意，上式有一个假设前提—— $X$  为有  $n$  个等概率值  $1/n$  的平均分布。否则，我们要把  $1/n$  替换成具体的概率值  $p_i$ 。不做特殊说明时，本书默认总体或样本取值都为等概率。

对于样本来说，随机变量  $X$  方差可以用连续分布的样本来估计：

$$\text{var}(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - E(X))^2 \quad (33)$$

对于数据矩阵  $\mathbf{X}$  而言，第  $j$  列数据  $\mathbf{x}_j$  的方差有几种不同表达方式：

$$\text{var}(X_j) = \text{var}(\mathbf{x}_j) = \sigma_j^2 = \sigma_{j,j} \quad (34)$$

## 中心化矩阵

利用中心化矩阵  $\mathbf{M}$ ,  $\sum_{i=1}^n (x_i - E(X))^2$  可以写成:

$$\sum_{i=1}^n (x_i - E(X))^2 = (\mathbf{M}\mathbf{x})^T \mathbf{M}\mathbf{x} = \mathbf{x}^T \mathbf{M}^T \mathbf{M}\mathbf{x} = \mathbf{x}^T \mathbf{M}\mathbf{x} \quad (35)$$

此外, 利用向量范数,  $\sum_{i=1}^n (x_i - E(X))^2$  还可以写成:

$$\sum_{i=1}^n (x_i - E(X))^2 = (\mathbf{x} - E(\mathbf{x}))^T (\mathbf{x} - E(\mathbf{x})) = \|\mathbf{x} - E(\mathbf{x})\|_2^2 \quad (36)$$

上式也用到了“广播原则”。

## 向量视角

图 10 中,  $\mathbf{x}$  在  $\mathbf{I}$  方向上向量投影为  $E(\mathbf{x})\mathbf{I}$ 。相当于  $\mathbf{x}$  被分解成  $E(\mathbf{x})\mathbf{I}$  和  $\mathbf{x} - E(\mathbf{x})\mathbf{I}$  两个向量分量。

$E(\mathbf{x})\mathbf{I}$  和  $\mathbf{I}$  平行, 而  $\mathbf{x} - E(\mathbf{x})\mathbf{I}$  和  $\mathbf{I}$  垂直。而向量  $\mathbf{x} - E(\mathbf{x})\mathbf{I}$  的模的平方就是 (36), 即:

$$\|\mathbf{x} - E(\mathbf{x})\mathbf{I}\|_2^2 = \sum_{i=1}^n (x_i - E(X))^2 \quad (37)$$

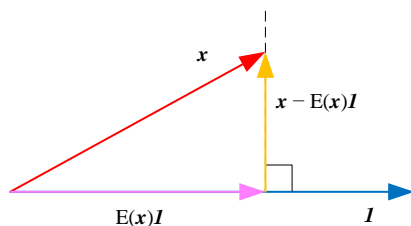


图 10. 投影角度看方差和标准差

## 鸢尾花数据

计算鸢尾花数据  $\mathbf{X}$  每一列标准差, 以向量方式表达:

$$\sigma_{\mathbf{X}} = \begin{bmatrix} 0.825 & 0.434 & 1.759 & 0.759 \\ \text{Sepal length} & \text{Sepal width} & \text{Petal length} & \text{Petal width} \end{bmatrix}^T \quad (38)$$

$X$  第三个特征，也就是花瓣长度  $X_3$  对应的标准差最大。图 11 所示为 KDE 估计得到的鸢尾花四个特征分布图。

➔ KDE 是核密度估计 (Kernel Density Estimation, KDE)，采用核函数拟合样本数据点，用来模拟样本数据在某一个特征上的分布情况。这是本系列丛书《概率统计》一册要讲解的话题。

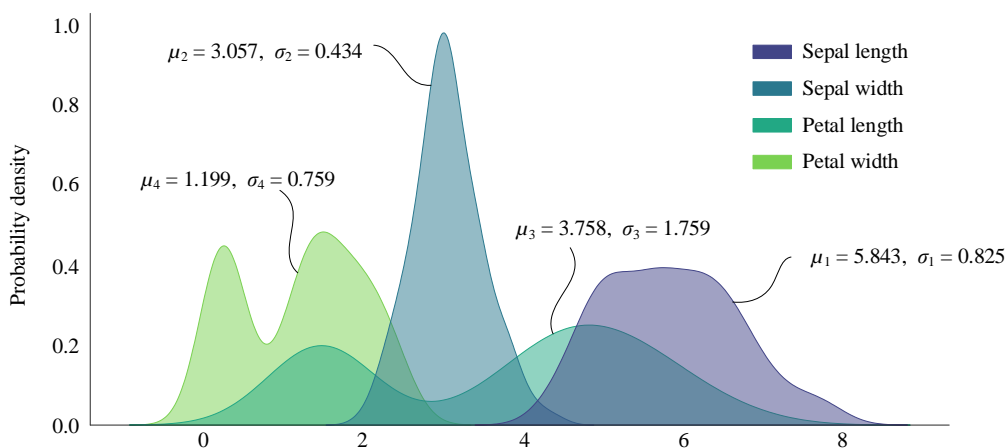


图 11. 鸢尾花数据四个特征上分布



Bk4\_Ch22\_01.py 中 Bk4\_Ch22\_01\_C 部分绘制图 11。

## 22.7 协方差和相关性系数

### 协方差

不考虑样本和总体的区别，列向量数据  $\mathbf{x}$  和  $\mathbf{y}$  协方差  $\text{cov}(\mathbf{x}, \mathbf{y})$  可以通过下式获得：

$$\begin{aligned} \text{cov}(\mathbf{x}, \mathbf{y}) &= \frac{(\mathbf{x} - \mathbf{E}(\mathbf{x})\mathbf{I})^T (\mathbf{y} - \mathbf{E}(\mathbf{y})\mathbf{I})}{n} = \frac{n\mathbf{x}^T \mathbf{y} - \mathbf{x}^T \mathbf{I} \mathbf{y}^T \mathbf{I}}{n^2} \\ &= \frac{\sum_{i=1}^n (x_i - \mathbf{E}(X))(y_i - \mathbf{E}(Y))}{n} = \frac{n \left( \sum_{i=1}^n x_i y_i \right) - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n^2} \end{aligned} \quad (39)$$

注意，上式同样有假设前提，即随机变量  $(X, Y)$  取到  $(x_i, y_i)$  的概率均为  $1/n$ 。

对于数据矩阵  $X$ ，列向量  $\mathbf{x}_i$  和  $\mathbf{x}_j$  的协方差有几种不同表达方式：

$$\text{cov}(X_i, X_j) = \text{cov}(\mathbf{x}_i, \mathbf{x}_j) = \rho_{i,j} \sigma_i \sigma_j = \sigma_{i,j} \quad (40)$$

## 中心化矩阵

利用中心化矩阵  $\mathbf{M}$ ,  $\sum_{i=1}^n (x_i - E(X))(y_i - E(Y))$  可以写成:

$$\sum_{i=1}^n (x_i - E(X))(y_i - E(Y)) = (\mathbf{M}\mathbf{x})^T \mathbf{M}\mathbf{y} = \mathbf{x}^T \mathbf{M}^T \mathbf{M}\mathbf{y} = \mathbf{x}^T \mathbf{M}\mathbf{y} \quad (41)$$

联合 (35) 和 (41), 下式成立:

$$\begin{bmatrix} \sum_{i=1}^n (x_i - E(X))^2 & \sum_{i=1}^n (x_i - E(X))(y_i - E(Y)) \\ \sum_{i=1}^n (y_i - E(Y))(x_i - E(X)) & \sum_{i=1}^n (y_i - E(Y))^2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T \mathbf{M}\mathbf{x} & \mathbf{x}^T \mathbf{M}\mathbf{y} \\ \mathbf{y}^T \mathbf{M}\mathbf{x} & \mathbf{y}^T \mathbf{M}\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^T \\ \mathbf{y}^T \end{bmatrix} \mathbf{M} \begin{bmatrix} \mathbf{x} & \mathbf{y} \end{bmatrix} \quad (42)$$

上式中, 协方差矩阵已经呼之欲出!

## 相关性系数

随机变量  $X$  和  $Y$  相关性系数的定义为:

$$\rho_{X,Y} = \text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\text{var}(Y)}} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (43)$$

相关性系数可以看做是随机变量  $z$  分数的协方差。

用向量内积形式来写, 列向量数据  $\mathbf{x}$  和  $\mathbf{y}$  相关性系数  $\text{corr}(\mathbf{x}, \mathbf{y})$  计算式如下:

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - E(\mathbf{x})) \cdot (\mathbf{y} - E(\mathbf{y}))}{\|\mathbf{x} - E(\mathbf{x})\| \|\mathbf{y} - E(\mathbf{y})\|} = \left( \frac{\mathbf{x} - E(\mathbf{x})}{\|\mathbf{x} - E(\mathbf{x})\|} \right) \cdot \left( \frac{\mathbf{y} - E(\mathbf{y})}{\|\mathbf{y} - E(\mathbf{y})\|} \right) \quad (44)$$

相信大家已经在上式中看到“平移”和“缩放”两步几何操作。上式把线性相关系数和向量内积联系起来。本书第 2 章介绍的余弦相似度 (cosine similarity) 也是通过两个向量的夹角的余弦值来度量它们之间的相似性:

$$\text{cosine similarity} = \cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \quad (45)$$

大家已经发现上两式在形式上高度相似。

## 向量内积、协方差



实际上，向量内积和协方差相似之处更多。比如，向量内积和协方差都满足交换律：

$$\begin{aligned} \mathbf{x} \cdot \mathbf{y} &= \mathbf{y} \cdot \mathbf{x} \\ \text{cov}(X, Y) &= \text{cov}(Y, X) \end{aligned} \quad (46)$$

向量的模类似标准差：

$$\begin{aligned} \|\mathbf{x}\| &= \sqrt{\mathbf{x} \cdot \mathbf{x}} \\ \sigma_X &= \sqrt{\text{var}(X)} = \sqrt{\text{cov}(X, X)} \end{aligned} \quad (47)$$

向量之间夹角余弦值类似线性相关性系数：

$$\begin{aligned} \cos \theta &= \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \\ \rho_{X,Y} = \text{corr}(X, Y) &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{E}((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \end{aligned} \quad (48)$$

(48) 可以分别整理成如下等式：

$$\begin{aligned} \mathbf{x} \cdot \mathbf{y} &= \cos \theta \|\mathbf{x}\| \|\mathbf{y}\| \\ \text{cov}(X, Y) &= \rho_{X,Y} \sigma_X \sigma_Y \end{aligned} \quad (49)$$

此外，余弦定理可以用在向量内积和协方差上：

$$\begin{aligned} \|\mathbf{x} + \mathbf{y}\|^2 &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 + 2\|\mathbf{x}\| \|\mathbf{y}\| \cos \theta \\ \text{var}(X + Y) &= \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y) \\ \sigma_{X+Y}^2 &= \sigma_X^2 + \sigma_Y^2 + 2\rho_{X,Y} \sigma_X \sigma_Y \\ \text{var}(aX + bY) &= a^2 \text{var}(X) + b^2 \text{var}(Y) + 2ab \text{cov}(X, Y) \end{aligned} \quad (50)$$

余弦的取值范围是  $[-1, 1]$ ，线性相关系数的取值范围也是  $[-1, 1]$ 。图 12 所示为余弦相似度和夹角  $\theta$  关系。

有了这种类比，下一章，我们将创造“标准差向量”，用向量视角解释质心、标准差、方差、协方差、协方差矩阵等统计描述。

**▲** 值得注意的是，统计中的方差和协方差运算都存在“中心化”，即去均值。也就是说，从几何角度来看，方差和协方差运算中都默认将“向量”起点移动到质心。

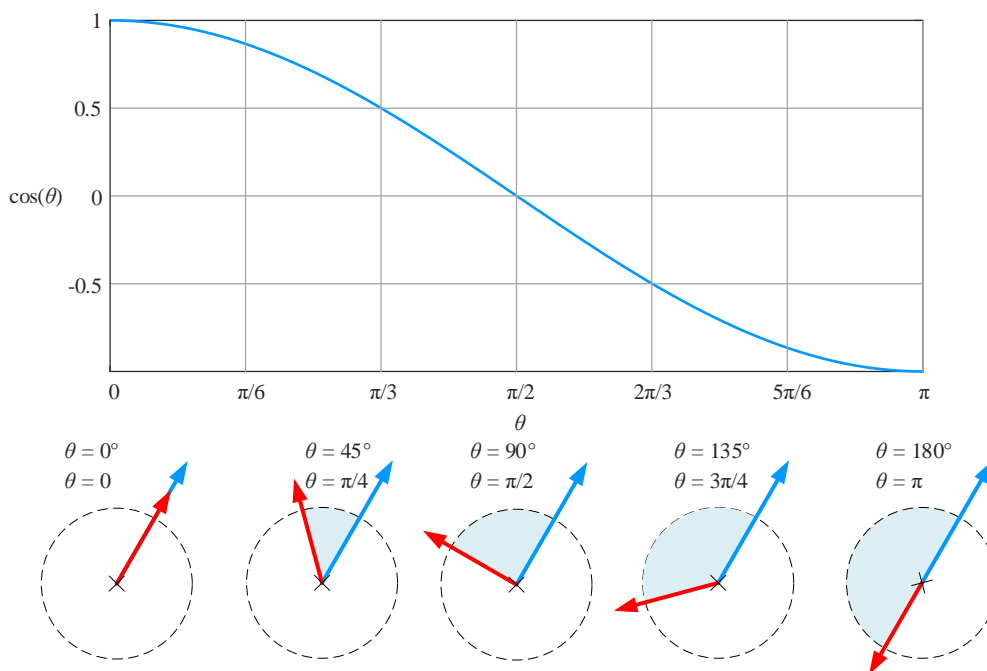


图 12. 余弦相似度

## 22.8 协方差矩阵和相关性系数矩阵

### 协方差矩阵

对于矩阵  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D]$  每两个列向量数据之间的协方差可以构造得到**协方差矩阵** (covariance matrix):

$$\mathbf{\Sigma} = \begin{bmatrix} \text{cov}(\mathbf{x}_1, \mathbf{x}_1) & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_1, \mathbf{x}_D) \\ \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & \text{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_2, \mathbf{x}_D) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(\mathbf{x}_D, \mathbf{x}_1) & \text{cov}(\mathbf{x}_D, \mathbf{x}_2) & \cdots & \text{cov}(\mathbf{x}_D, \mathbf{x}_D) \end{bmatrix} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D,1} & \sigma_{D,2} & \cdots & \sigma_{D,D} \end{bmatrix} \quad (51)$$

很明显协方差矩阵是对称矩阵。协方差矩阵又叫方差-协方差矩阵，这是因为  $\mathbf{\Sigma}$  对角线元素均为方差，其余元素为协方差。

样本协方差矩阵  $\mathbf{\Sigma}$  则可以用数据矩阵  $\mathbf{X}$  计算得到：

$$\mathbf{\Sigma} = \frac{\left( \underbrace{\mathbf{X} - \mathbf{E}(\mathbf{X})}_{\text{Centered}} \right)^T \left( \underbrace{\mathbf{X} - \mathbf{E}(\mathbf{X})}_{\text{Centered}} \right)}{n-1} \quad (52)$$

对于总体，分母则改为  $n$ 。特别地，如果  $n$  足够大， $n$  和  $n-1$  对计算影响可以忽略不计。

用中心化数据  $\mathbf{X}_c$  代替  $\mathbf{X} - \mathbf{E}(\mathbf{X})$ ，(52) 可以写成：

$$\boldsymbol{\Sigma} = \frac{\overbrace{\mathbf{X}_c^T \mathbf{X}_c}^{\text{Gram matrix}}}{n-1} \quad (53)$$

相信大家已经在上式中看到了格拉姆矩阵。这也就是说，协方差矩阵  $\boldsymbol{\Sigma}$  某种程度上就是  $\mathbf{X}_c$  的格拉姆矩阵。

## 特征值分解

由于协方差矩阵为对称矩阵，对  $\boldsymbol{\Sigma}$  进行特征值分解，得到：

$$\boldsymbol{\Sigma} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \quad (54)$$

得知协方差矩阵为对称矩阵，不知道大家是否立刻想到本书第 20 章介绍的二次型，将  $\boldsymbol{\Sigma}$  写成二次型  $\mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x}$ 。将 (54) 代入  $\mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x}$ ，得到：

$$\begin{aligned} \mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x} &= \mathbf{x}^T \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \mathbf{x} = \left( \mathbf{V}^T \mathbf{x} \right)^T \boldsymbol{\Lambda} \left( \mathbf{V}^T \mathbf{x} \right) = \mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y} \\ &= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_D y_D^2 = \sum_{j=1}^D \lambda_j y_j^2 \end{aligned} \quad (55)$$

从几何角度来看， $\mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y}$  就是正椭圆，这意味着  $\mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x}$  为旋转椭圆。

特别地，当  $D=2$  时， $\mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x}$  代表旋转椭圆：

$$\mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \sigma_{1,1} x_1^2 + (\sigma_{1,2} + \sigma_{2,1}) x_1 x_2 + \sigma_{2,2} x_2^2 \quad (56)$$

$\mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y}$  为正椭圆：

$$\mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \lambda_1 y_1^2 + \lambda_2 y_2^2 \quad (57)$$

如图 13 所示，正是 (54) 中的  $\mathbf{V}$  完成正椭圆到旋转椭圆的“旋转”。如果大家对于几何变换细节感到陌生的话，请回顾本书第 14、20 章。

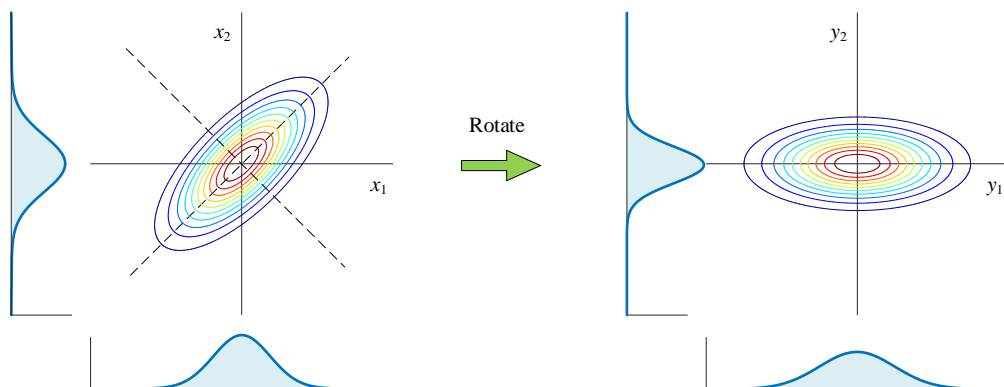


图 13. 旋转椭圆到正椭圆

## 相关性系数矩阵

**相关性系数矩阵** (correlation matrix)  $\mathbf{P}$  定义为：

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix} \quad (58)$$

$\mathbf{P}$  和  $\Sigma$  的关系为：

$$\Sigma = \mathbf{S} \mathbf{P} \mathbf{S}^T \quad (59)$$

$\mathbf{S}$  就是 (26) 定义的缩放矩阵， $\mathbf{S}$  是个对角方阵。

再进一步，(58) 可以写成：

$$\mathbf{P} = \frac{\left( \underbrace{(\mathbf{X} - \mathbf{E}(\mathbf{X})) \mathbf{S}^{-1}}_{\text{Translate + Scale}} \right)^T \left( \underbrace{(\mathbf{X} - \mathbf{E}(\mathbf{X})) \mathbf{S}^{-1}}_{\text{Translate + Scale}} \right)}{n-1} \quad (60)$$

我们可以在上式中看到“平移”、“缩放”两步操作。

同时，我们在 (60) 中看到了 (25) 定义的 z 分数矩阵  $\mathbf{Z}_X$ 。因此，(60) 可以写成：

$$\mathbf{P} = \frac{\mathbf{Z}_X^T \mathbf{Z}_X}{n-1} \quad (61)$$

相关性系数矩阵  $\mathbf{P}$  可以看做  $\mathbf{Z}_X$  的协方差矩阵。也就是说， $\mathbf{P}$  相当于的格拉姆矩阵  $\mathbf{Z}_X$ 。准确地说， $\mathbf{Z}_X$  的格拉姆矩阵为  $\mathbf{Z}_X^T \mathbf{Z}_X = (n-1) \mathbf{P}$ 。

## 鸢尾花数据集

对于鸢尾花数据，它的协方差矩阵  $\Sigma$  为：

$$\Sigma = \begin{bmatrix} 0.686 & -0.042 & 1.274 & 0.516 \\ -0.042 & 0.190 & -0.330 & -0.122 \\ 1.274 & -0.330 & 3.116 & 1.296 \\ 0.516 & -0.122 & 1.296 & 0.581 \end{bmatrix} \begin{matrix} \leftarrow \text{Sepal length} \\ \leftarrow \text{Sepal width} \\ \leftarrow \text{Petal length} \\ \leftarrow \text{Petal width} \end{matrix} \quad (62)$$

鸢尾花数据的相关性系数矩阵  $P$  为：

$$P = \begin{bmatrix} 1.000 & -0.118 & 0.872 & 0.818 \\ -0.118 & 1.000 & -0.428 & -0.366 \\ 0.872 & -0.428 & 1.000 & 0.963 \\ 0.818 & -0.366 & 0.963 & 1.000 \end{bmatrix} \begin{matrix} \leftarrow \text{Sepal length} \\ \leftarrow \text{Sepal width} \\ \leftarrow \text{Petal length} \\ \leftarrow \text{Petal width} \end{matrix} \quad (63)$$

图 14 所示为  $\Sigma$  和  $P$  的热图。观察相关性系数矩阵  $P$ ，可以发现花萼长度和花萼宽度线性负相关，花瓣长度和花萼宽度线性负相关，花瓣宽度和花萼宽度线性负相关。当然，鸢尾花数据集样本数量有限，通过样本数据得出的结论还不足以推而广之。

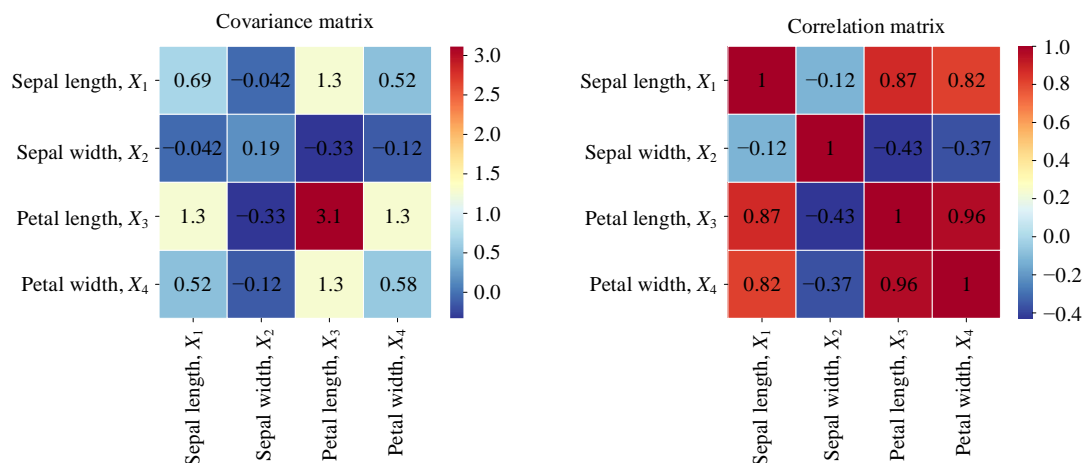


图 14. 协方差矩阵和相关性系数矩阵热图

本系列丛书《概率统计》会建立协方差矩阵和椭圆的密切关系。图 15 便来自《概率统计》，图中我们可以通过椭圆的大小和旋转角度了解不同特征标准差，以及不同特征之间的相关性这样的重要信息。

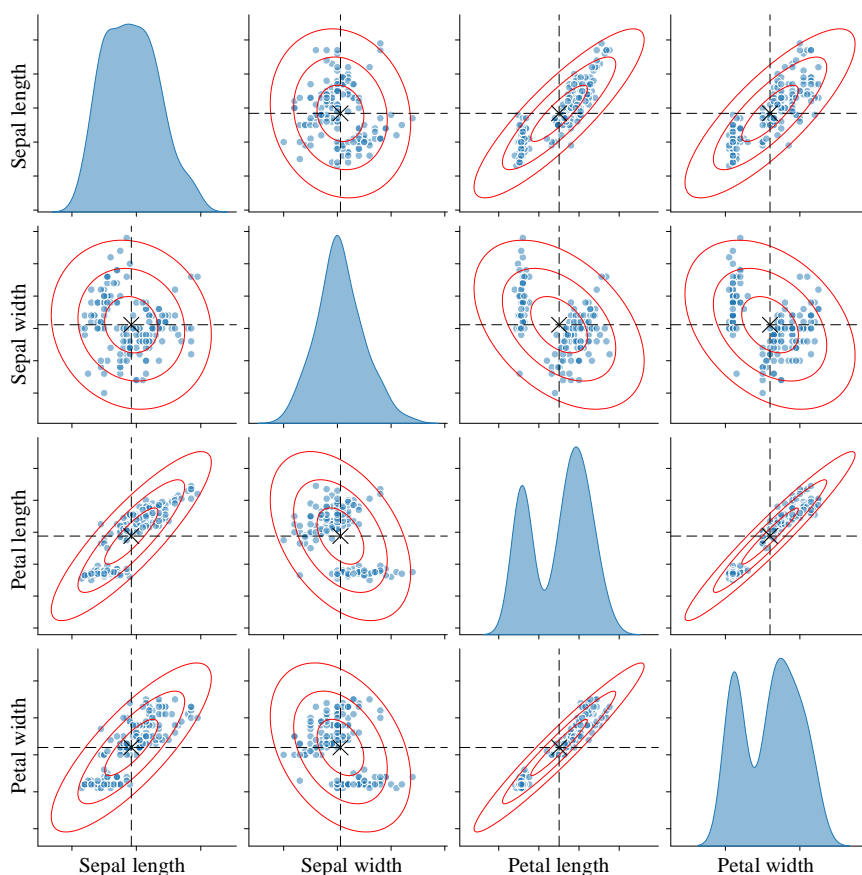


图 15. 协方差矩阵和椭圆的关系

如前文所述，鸢尾花数据分为三类。标签为  $C_k$  样本数据也对应自身协方差矩阵  $\Sigma_k$  (如图 16) 和相关性系数矩阵  $\mathbf{P}_k$  (如图 17)。图 18 也是来自本系列丛书《概率统计》一册，图中绘制椭圆时考虑鸢尾花分类。这些旋转椭圆的中心就是簇质心，椭圆本身代表簇协方差矩阵。

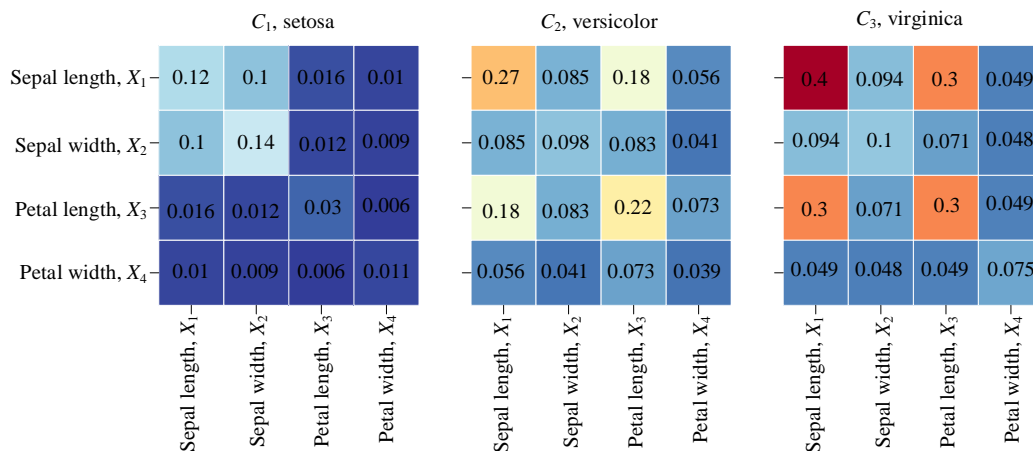


图 16. 协方差矩阵热图，考虑分类

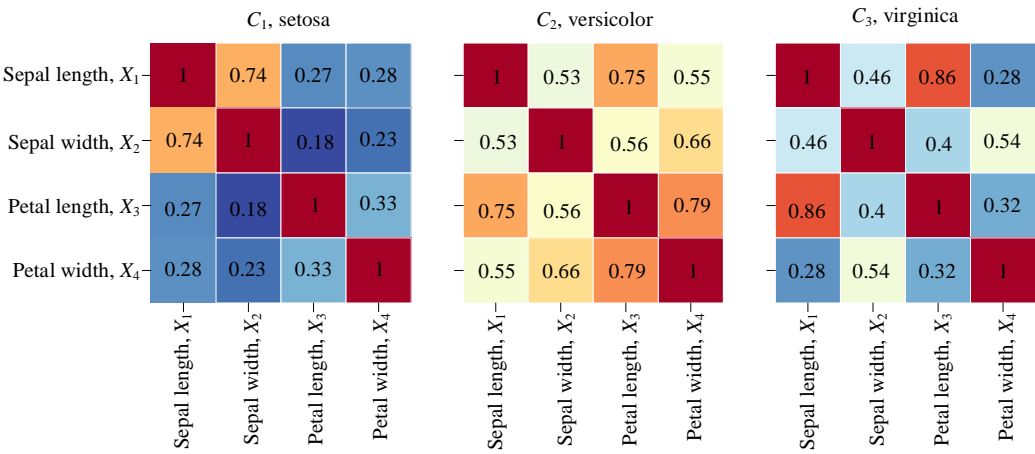


图 17. 相关性系数矩阵热图，考虑分类

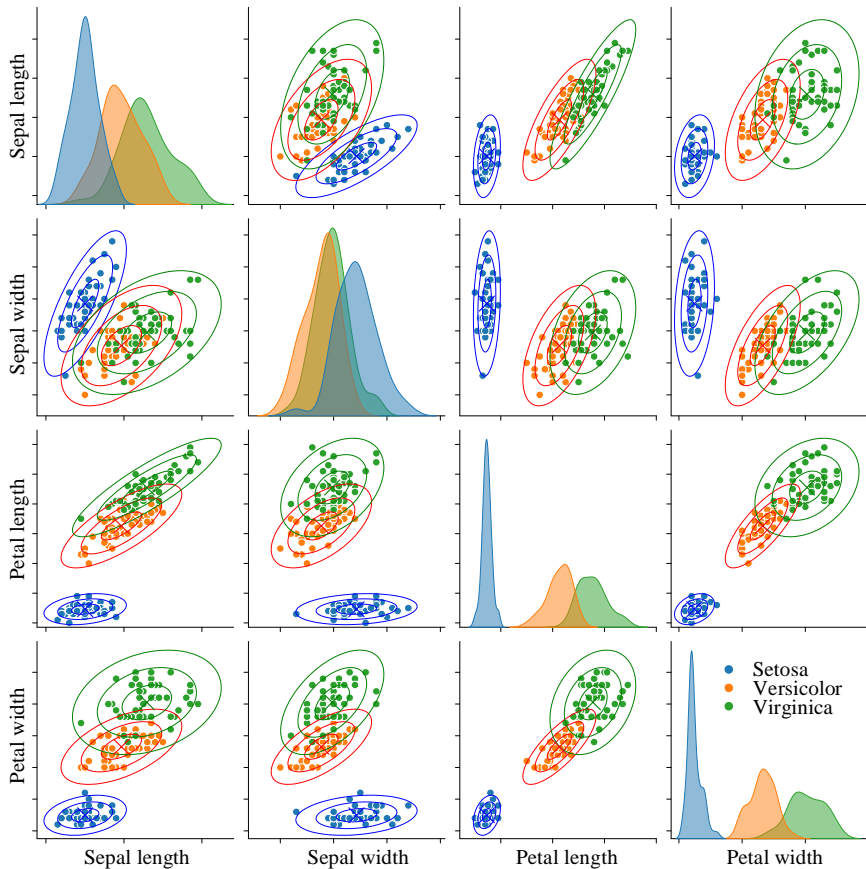


图 18. 协方差矩阵和椭圆的关系，考虑分类



Bk4\_Ch22\_01.py 中 Bk4\_Ch22\_01\_D 部分绘制图 14、图 16、图 17 这几幅热图。



本章从线性代数运算视角回顾、梳理统计学中一些重要的概念。希望大家学完本章后，能够轻松建立数据、矩阵、向量、统计之间的联系。

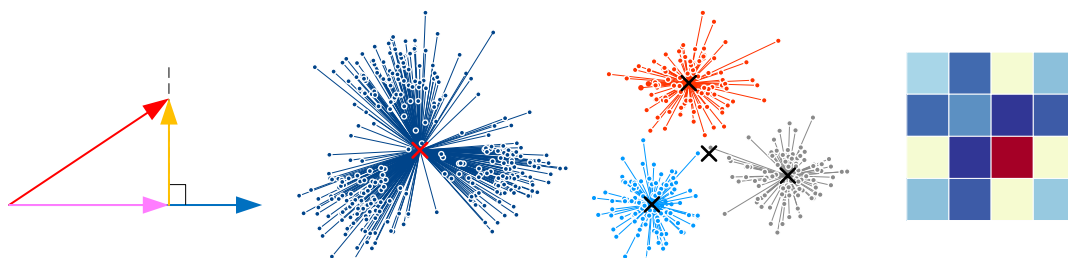


图 19. 总结本章重要内容的四幅图

本章介绍了两种和原始数据  $\mathbf{X}$  形状相同的数据矩阵——中心化数据矩阵  $\mathbf{X}_c$ 、标准化数据矩阵  $\mathbf{Z}_x$ 。请大家注意它们三者区分和联系。并且能从几何变换视角理解运算过程。

质心和协方差矩阵在后续众多数据科学、机器学习算法中扮演重要角色。此外，请大家务必注意协方差矩阵和椭圆之间的千丝万缕的联系。本系列丛书《概率统计》将从不同角度讲解如何利用椭圆更好地理解高斯分布、条件概率、线性回归、主成分分析等数学工具。

下一章正式进入本书收关之旅——数据三部曲。



推荐大家阅读多元统计方面的一本经典，Richard A. Johnson 和 Dean W. Wichern 合著的 *Applied Multivariate Statistical Analysis*。清华大学出版社翻译出版了这部作品，书名为《实用多元统计分析》。