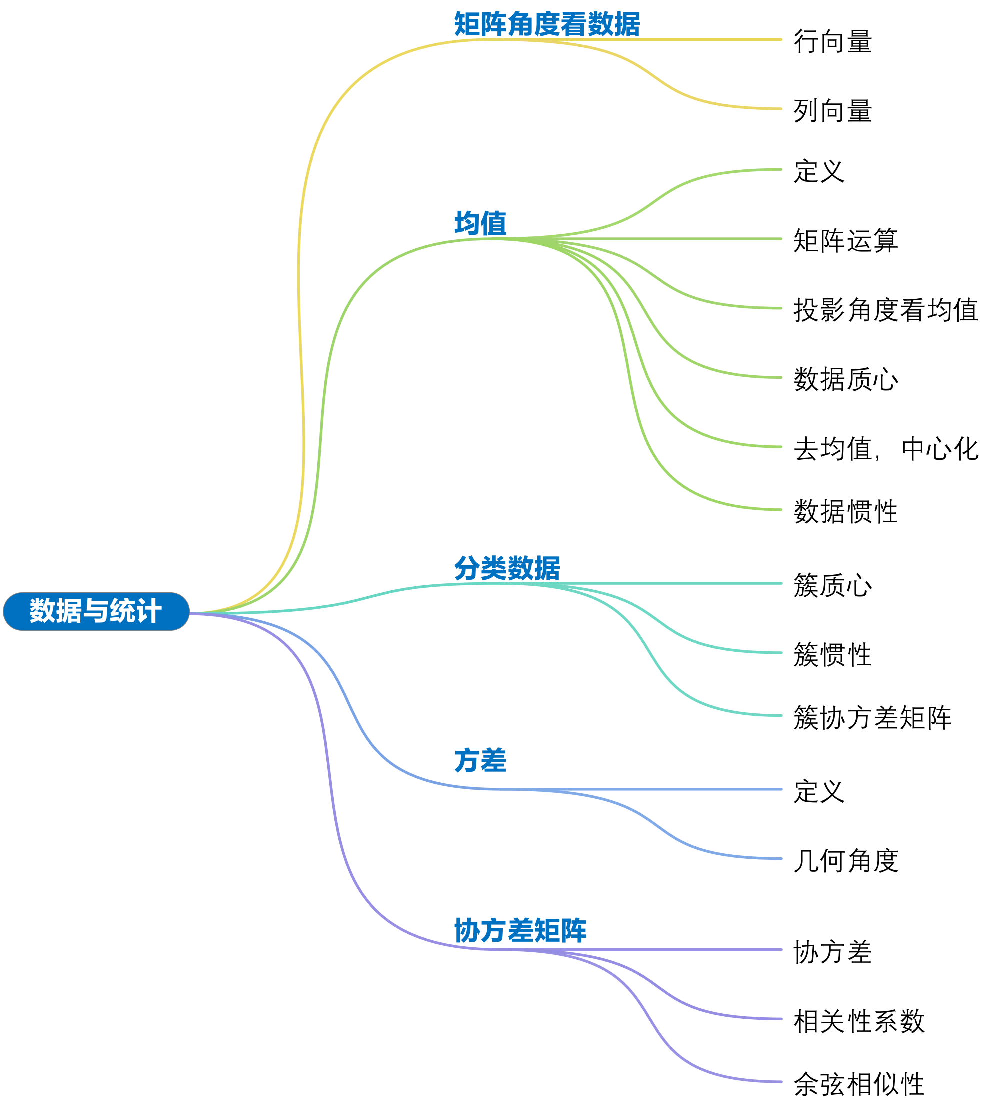
22 Statistics Meet Linear Algebra 数据与统计

有数据的地方，怎么能没有统计

> 毫无争议的是，人类无法准确地地判断事物的真伪，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


数据与统计
矩阵角度看数据
行向量
列向量
均值
定义
矩阵运算
投影角度看均值
数据质心
去均值，中心化
数据惯性
分类数据
簇质心
簇惯性
簇协方差矩阵
方差
定义
几何角度
协方差矩阵
协方差
相关性系数
余弦相似性

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 $\boldsymbol{X}$ 的形状为 150 × 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 $\boldsymbol{X}$ 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。

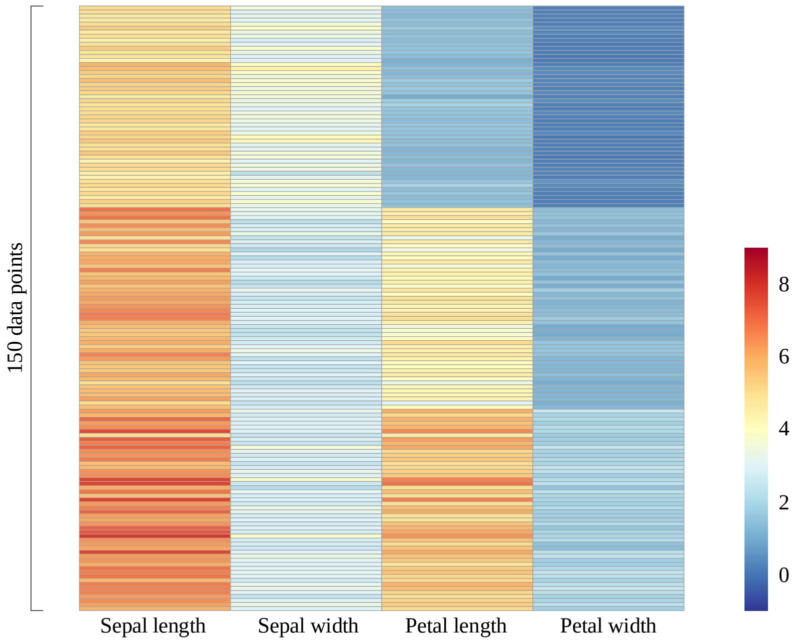


图 1. 鸢尾花数据，原始数据矩阵 $\boldsymbol{X}$，单位为厘米 (cm)

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

$$\boldsymbol{x}_j = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix} \tag{1}$$

列向量 $\boldsymbol{x}_j$ 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) $\mathrm{E}(X_j)$：

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

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

$$\mathrm{E}\left(\boldsymbol{x}_j\right) = \mathrm{E}\left(X_j\right) = \mu_j = \frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}}{n} = \frac{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{x}_j}{n} = \frac{\boldsymbol{x}_j \cdot \boldsymbol{1}}{n} = \frac{\boldsymbol{1} \cdot \boldsymbol{x}_j}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \tag{3}$$

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

(3) 左乘 n 可以得到如下等式：

$$n\mu_j = n\mathrm{E}\left(\boldsymbol{x}_j\right) = \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1} = \boldsymbol{1}^{\mathrm{T}}\boldsymbol{x}_j = \boldsymbol{x}_j \cdot \boldsymbol{1} = \boldsymbol{1} \cdot \boldsymbol{x}_j \tag{4}$$

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

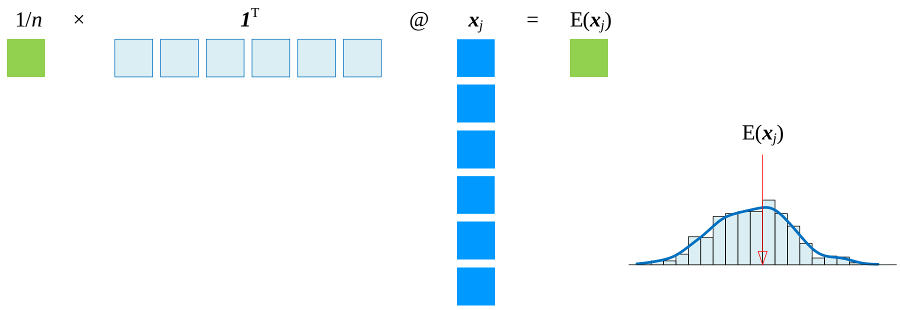


图 2. 计算 $\boldsymbol{x}_j$ 期望值/均值

利用矩阵运算分别得到鸢尾花的四个特征的期望值：

$$\begin{cases} \mathrm{E}\left(\boldsymbol{x}_1\right)=\mu_1=\underbrace{5.843}_{\text{Sepal length}} \\ \mathrm{E}\left(\boldsymbol{x}_2\right)=\mu_2=\underbrace{3.057}_{\text{Sepal width}} \\ \mathrm{E}\left(\boldsymbol{x}_3\right)=\mu_3=\underbrace{3.758}_{\text{Petal length}} \\ \mathrm{E}\left(\boldsymbol{x}_4\right)=\mu_4=\underbrace{1.199}_{\text{Petal width}} \end{cases} \tag{5}$$

向量视角

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

如图 3 所示，$\mathrm{E}(\boldsymbol{x}_j)$ 是一个标量，而向量 $\mathrm{E}(\boldsymbol{x}_j)\boldsymbol{1}$ 相当于向量 $\boldsymbol{x}_j$ 在 $\boldsymbol{1}$ 方向上投影的向量投影结果：

$$\mathrm{E}\left(\boldsymbol{x}_j\right)\boldsymbol{1}=\operatorname{proj}_{\boldsymbol{1}}\left(\boldsymbol{x}_j\right)=\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}}{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{1}}\boldsymbol{1}=\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}}{n}\boldsymbol{1} \tag{6}$$

⚠ 再次注意，$\mathrm{E}(\boldsymbol{x}_j)$ 为标量；$\mathrm{E}(\boldsymbol{x}_j)\boldsymbol{1}$ 为向量，和 $\boldsymbol{1}$ 平行。

图 3 中，$\boldsymbol{1}$ 方向上解释了 $\boldsymbol{x}_j$ 中 $\mathrm{E}(\boldsymbol{x}_j)\boldsymbol{1}$ 这部分分量，没有被解释的向量分量为：

$$\boldsymbol{x}_j-\operatorname{proj}_{\boldsymbol{1}}\left(\boldsymbol{x}_j\right)=\boldsymbol{x}_j-\mathrm{E}\left(\boldsymbol{x}_j\right)\boldsymbol{1} \tag{7}$$

(7) 这部分垂直于 $\boldsymbol{1}$，也就是说：

$$\boldsymbol{1}^{\mathrm{T}}\left(\boldsymbol{x}_j-\operatorname{proj}_{\boldsymbol{1}}\left(\boldsymbol{x}_j\right)\right)=\boldsymbol{1}^{\mathrm{T}}\left(\boldsymbol{x}_j-\overbrace{\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}}{n}}^{\text{Scalar}}\boldsymbol{1}\right)=\boldsymbol{1}^{\mathrm{T}}\boldsymbol{x}_j-\frac{\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}}{n}\underbrace{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{1}}_{n}=\boldsymbol{1}^{\mathrm{T}}\boldsymbol{x}_j-\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}=0 \tag{8}$$

注意，上式中 $\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}$ 为标量，因此 $\boldsymbol{1}^{\mathrm{T}}\left(\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}\right)\boldsymbol{1}=\left(\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{1}\right)\boldsymbol{1}^{\mathrm{T}}\boldsymbol{1}$。均值作为一个统计量，它能解释列向量 $\boldsymbol{x}_j$ 一部分特征。$\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_j)\boldsymbol{1}$ 将在标准差 (方差平方根) 中加以解释。

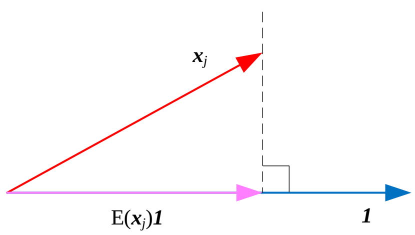


图 3. 投影角度看期望值

两个极端例子

如果 x_j 所有元素均相同，比如全都是 k，那么 x_j 可以写成：

$$\boldsymbol{x}_j = \begin{bmatrix} k \\ k \\ \vdots \\ k \end{bmatrix} = k \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = k\boldsymbol{1} \tag{9}$$

这种情况，$\boldsymbol{x}_j$ 和 $\boldsymbol{1}$ 共线。

再举个相反的例子，如果 $\boldsymbol{x}_j$ 和 $\boldsymbol{1}$ 垂直，

$$\boldsymbol{1}^{\mathrm{T}} \boldsymbol{x}_j = 0 \tag{10}$$

也就是意味着 $\mathrm{E}(\boldsymbol{x}_j) = 0$。也就是说，$\boldsymbol{x}_j$ 在 $\boldsymbol{1}$ 方向的标量投影为 0。

对于最小二乘法线性回归，$\boldsymbol{x}_j - \mathrm{E}(\boldsymbol{x}_j)\boldsymbol{1}$ 垂直于 $\boldsymbol{1}$ 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 $\boldsymbol{X}$ 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 $\boldsymbol{X}$ 的质心位置。

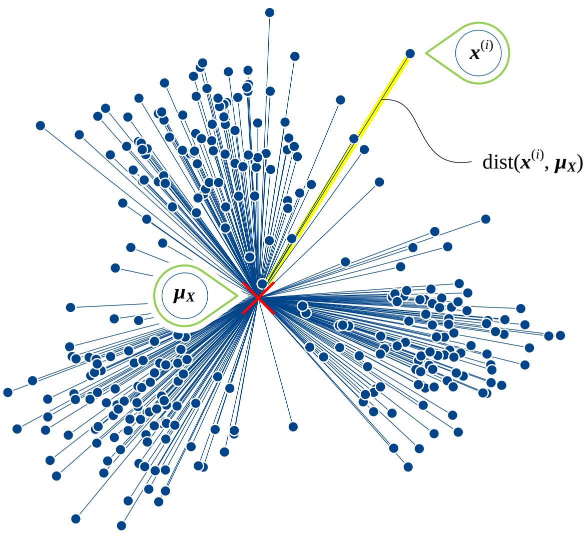


图 4. 平面上数据矩阵 $\boldsymbol{X}$ 质心位置

列向量

$\boldsymbol{X}$ 样本数据的质心 $\boldsymbol{\mu}_X$ 定义如下：

$$\boldsymbol{\mu}_X = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} = \begin{bmatrix} \mathrm{E}(\boldsymbol{x}_1) \\ \mathrm{E}(\boldsymbol{x}_2) \\ \vdots \\ \mathrm{E}(\boldsymbol{x}_D) \end{bmatrix} \tag{11}$$

⚠ 注意，为了方便运算，$\boldsymbol{\mu}_X$ 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 $\boldsymbol{\mu}_X$。比如，多元高斯分布的概率密度函数：

$$f_X(\boldsymbol{x}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_X)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_X)\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \tag{12}$$

上式中，几何角度来看，$\boldsymbol{x} - \boldsymbol{\mu}_X$ 相当于“平移”，$\boldsymbol{\Sigma}^{-1}$ 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，$\boldsymbol{\mu}_X$ 可以通过如下矩阵运算获得：

$$\boldsymbol{\mu}_X = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} = \frac{\left(\boldsymbol{1}^{\mathrm{T}} \boldsymbol{X}\right)^{\mathrm{T}}}{n} = \frac{\boldsymbol{X}^{\mathrm{T}} \boldsymbol{1}}{n} \tag{13}$$

其中，样本数据矩阵 $\boldsymbol{X}$ 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：

$$\begin{cases} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{1} = n\boldsymbol{\mu}_X \\ \boldsymbol{1}^{\mathrm{T}} \boldsymbol{X} = n(\boldsymbol{\mu}_X)^{\mathrm{T}} \end{cases} \tag{14}$$

举个例子，鸢尾花数据质心位置：

$$\boldsymbol{\mu}_X = \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \tag{15}$$

本书第 5 章讲过上述内容。

行向量

为了区分，丛书特别定义 $\mathrm{E}(\boldsymbol{X})$ 为行向量，即：

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{X}) &= \left[\mathrm{E}(\boldsymbol{x}_1) \quad \mathrm{E}(\boldsymbol{x}_2) \quad \cdots \quad \mathrm{E}(\boldsymbol{x}_D)\right] \\
&= \left[\mu_1 \quad \mu_2 \quad \cdots \quad \mu_D\right] \\
&= \left(\boldsymbol{\mu}_X\right)^{\mathrm{T}} = \frac{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{X}}{n}
\end{aligned}
\tag{16}
$$

整理 (16)，可以得到：

$$
\boldsymbol{1}^{\mathrm{T}}\boldsymbol{X} = n\mathrm{E}(\boldsymbol{X}) \tag{17}
$$

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

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

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

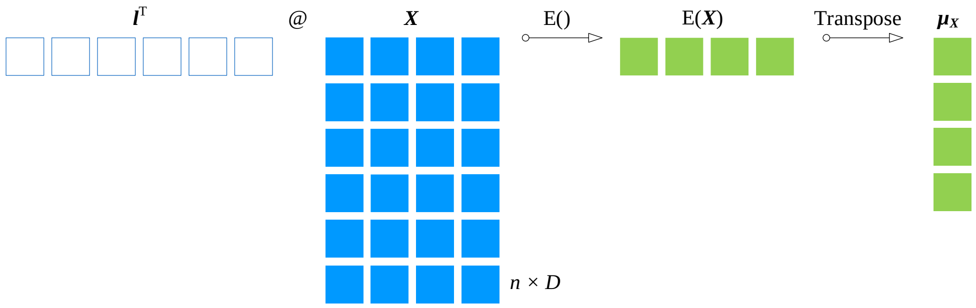


图 5. 计算 $\boldsymbol{X}$ 样本数据的质心 $\boldsymbol{\mu}_X$

22.4 中心化：平移

中心化、去均值

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

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

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

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

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

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

令：

$$\boxed{M = I - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}} \tag{20}$$

本章后文称 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 X，即 X 减去行向量 $\mathrm{E}(X)$ 得到 X_c：

$$X_c = X - \mathrm{E}(X) \tag{21}$$

中心化后，数据 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 M，下面我们简单分析一下这个特殊矩阵。

将 M 展开得到：

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

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

矩阵 M 为幂等矩阵，即满足：

$$MM = M \tag{23}$$

将 (20) 代入上式，展开整理：

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

我们在后文还会用到 M 这个中心化矩阵。

(24) 中所有全 1 列向量 $\mathbf{1}$ 等长，形状均为 $n \times 1$。因此 $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ 结果为 $n \times n$ 方阵，矩阵中每个元素都是 1。而 $\mathbf{1}^{\mathrm{T}}\mathbf{1}$ 结果为标量 n。我们也会在很多运算中看到 $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ 中两个 $\mathbf{1}$ 长度不同。此时，$\mathbf{1}\mathbf{1}^{\mathrm{T}}$ 结果为长方阵。此外，(24) 中两个单位矩阵 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 $IA_{m\times n}I = A_{m\times n}$ 这个等式左右的单位矩阵形状显然不同，左边 I 形状为 $m \times m$，右边 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 $\boldsymbol{X}_c$ 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：

$$\boldsymbol{Z}_X = \boldsymbol{X}_c \boldsymbol{S}^{-1} = \left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)\boldsymbol{S}^{-1} \tag{25}$$

其中，缩放矩阵 $\boldsymbol{S}$ 为：

$$\boldsymbol{S} = \mathrm{diag}\left(\mathrm{diag}\left(\boldsymbol{\Sigma}\right)\right)^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix} \tag{26}$$

其中，里层 diag() 提取对角线元素，结果为向量；外层 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：

$$\mathrm{SSD}\left(\boldsymbol{X}\right) = \sum_{i=1}^{n} \mathrm{dist}\left(\boldsymbol{x}^{(i)}, \mathrm{E}\left(\boldsymbol{X}\right)\right)^2 = \sum_{i=1}^{n} \left\|\boldsymbol{x}^{(i)} - \mathrm{E}\left(\boldsymbol{X}\right)\right\|_2^2 = \sum_{i=1}^{n} \left\|\boldsymbol{x}^{(i)\mathrm{T}} - \boldsymbol{\mu}_X\right\|_2^2 \tag{27}$$

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

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

$$\mathrm{SSD}\left(\boldsymbol{X}\right) = \mathrm{trace}\left(\boldsymbol{X}_c^{\mathrm{T}} \boldsymbol{X}_c\right) = \mathrm{trace}\left(\left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)^{\mathrm{T}}\left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)\right) \tag{28}$$

`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。

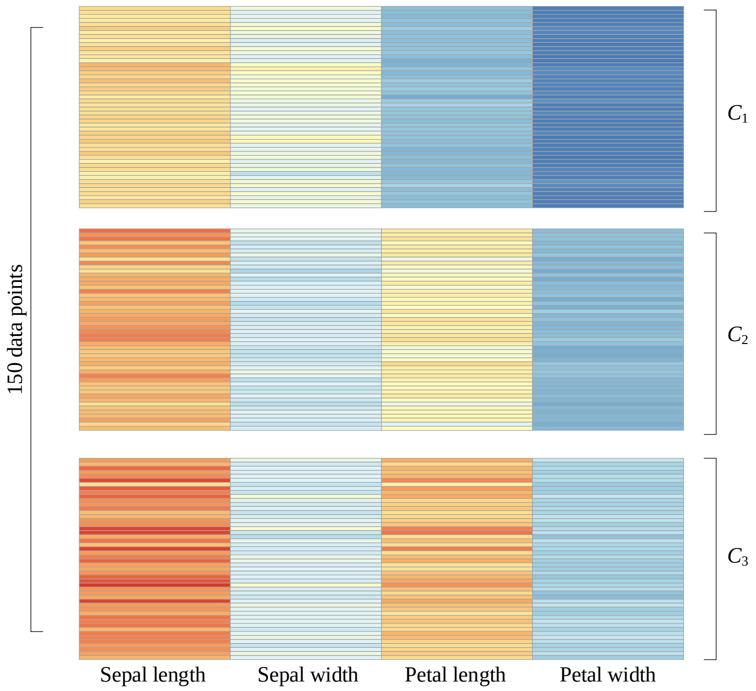


图 7. 鸢尾花数据分为三类

簇质心

类似 $\boldsymbol{\mu}_X$，任意一类标签为 C_k 样本数据的簇质心 $\boldsymbol{\mu}_k$，定义如下：

$$\boldsymbol{\mu}_k = \frac{1}{\operatorname{count}(C_k)} \sum_{i \in C_k} \boldsymbol{x}^{(i)\mathrm{T}} \tag{29}$$

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 $\boldsymbol{x}^{(i)}$ ($i \in C_k$)，求其各个特征平均值，构造成一个新的列向量 $\boldsymbol{\mu}_k$。图 8 所示为样本数据质心 $\boldsymbol{\mu}_X$，和三个不同标签数据各自的簇质心 $\boldsymbol{\mu}_1$、$\boldsymbol{\mu}_2$ 和 $\boldsymbol{\mu}_3$ 之间的关系。

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

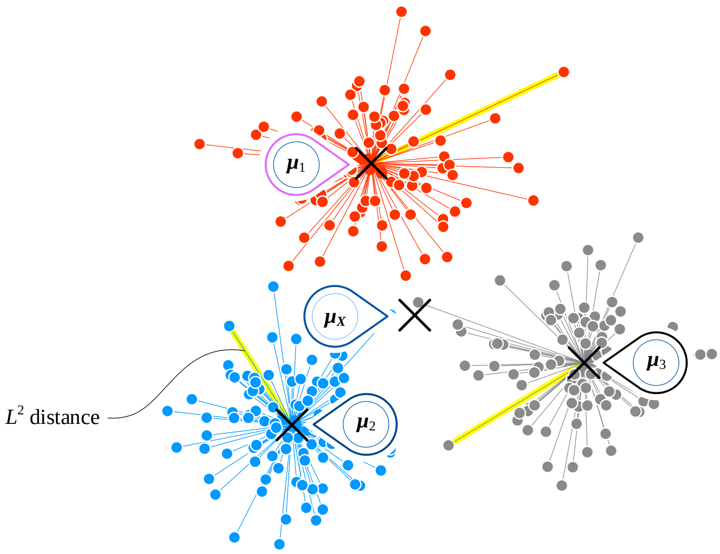


图 8. 样本数据质心 $\boldsymbol{\mu}_X$，和三类数据各自的质心 $\boldsymbol{\mu}_1$、$\boldsymbol{\mu}_2$ 和 $\boldsymbol{\mu}_3$

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：$\{(\boldsymbol{x}^{(2)}, y^{(2)} = C_1), (\boldsymbol{x}^{(5)}, y^{(5)} = C_1), (\boldsymbol{x}^{(6)}, y^{(6)} = C_1), (\boldsymbol{x}^{(9)}, y^{(9)} = C_1)\}$。

数据点有两特征，具体坐标值如下：

$$\boldsymbol{x}^{(2)} = \begin{bmatrix} 2 & 3 \end{bmatrix}, \quad \boldsymbol{x}^{(5)} = \begin{bmatrix} 3 & 1 \end{bmatrix}, \quad \boldsymbol{x}^{(6)} = \begin{bmatrix} -2 & 2 \end{bmatrix}, \quad \boldsymbol{x}^{(9)} = \begin{bmatrix} 1 & 6 \end{bmatrix} \tag{30}$$

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

$$\begin{aligned} \boldsymbol{\mu}_{C1} &= \frac{1}{\text{count}(C_1)} \sum_{i \in C_1} \boldsymbol{x}^{(i)\text{T}} = \frac{1}{\text{count}(C_1)} \left(\boldsymbol{x}^{(2)\text{T}} + \boldsymbol{x}^{(5)^\text{T}} + \boldsymbol{x}^{(6)^\text{T}} + \boldsymbol{x}^{(9)^\text{T}} \right) \\ &= \frac{1}{4} \left(\begin{bmatrix} 2 & 3 \end{bmatrix}^\text{T} + \begin{bmatrix} 3 & 1 \end{bmatrix}^\text{T} + \begin{bmatrix} -2 & 2 \end{bmatrix}^\text{T} + \begin{bmatrix} 1 & 6 \end{bmatrix}^\text{T} \right) = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \end{aligned} \tag{31}$$

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 × 代表不同标签鸢尾花的簇质心位置。

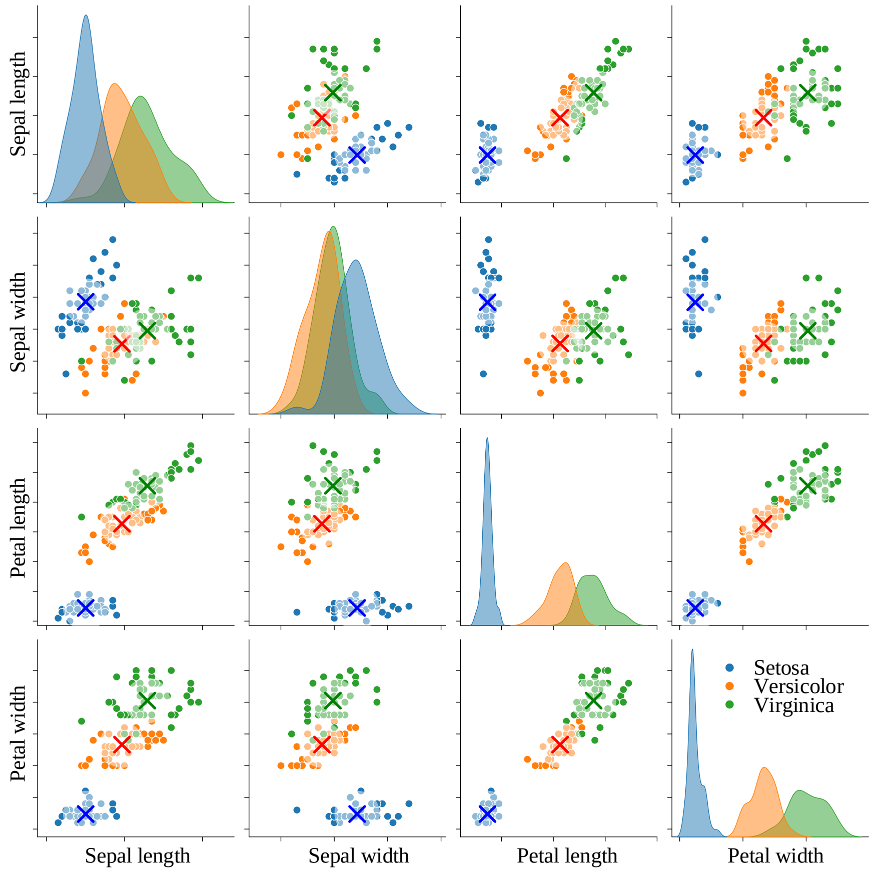


图 9. 鸢尾花数据簇质心位置

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：

$$\operatorname{var}(X)=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mathrm{E}(X)\right)^2 \tag{32}$$

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：

$$\operatorname{var}(X)=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\mathrm{E}(X)\right)^2 \tag{33}$$

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

$$\operatorname{var}(X_j)=\operatorname{var}(\boldsymbol{x}_j)=\sigma_j^2=\sigma_{j,j} \tag{34}$$

中心化矩阵

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

$$\sum_{i=1}^{n}\left(x_i-\mathrm{E}(X)\right)^2=(\boldsymbol{M}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} \tag{35}$$

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

$$\sum_{i=1}^{n}\left(x_i-\mathrm{E}(X)\right)^2=\left(\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\right)^{\mathrm{T}}\left(\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\right)=\left\|\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\right\|_2^2 \tag{36}$$

上式也用到了“广播原则”。

向量视角

图 10 中，$\boldsymbol{x}$ 在 $\boldsymbol{1}$ 方向上向量投影为 $\mathrm{E}(\boldsymbol{x})\boldsymbol{1}$。相当于 $\boldsymbol{x}$ 被分解成 $\mathrm{E}(\boldsymbol{x})\boldsymbol{1}$ 和 $\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\boldsymbol{1}$ 两个向量分量。

$\mathrm{E}(\boldsymbol{x})\boldsymbol{1}$ 和 $\boldsymbol{1}$ 平行，而 $\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\boldsymbol{1}$ 和 $\boldsymbol{1}$ 垂直。而向量 $\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\boldsymbol{1}$ 的模的平方就是 (36)，即：

$$\left\|\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\boldsymbol{1}\right\|_2^2=\sum_{i=1}^{n}\left(x_i-\mathrm{E}(X)\right)^2 \tag{37}$$

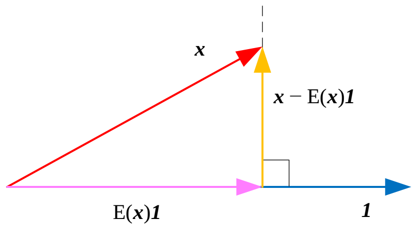


图 10. 投影角度看方差和标准差

鸢尾花数据

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

$$\boldsymbol{\sigma}_X=\begin{bmatrix}\underset{\text{Sepal length}}{0.825} & \underset{\text{Sepal width}}{0.434} & \underset{\text{Petal length}}{1.759} & \underset{\text{Petal width}}{0.759}\end{bmatrix}^{\mathrm{T}} \tag{38}$$

$\boldsymbol{X}$ 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。

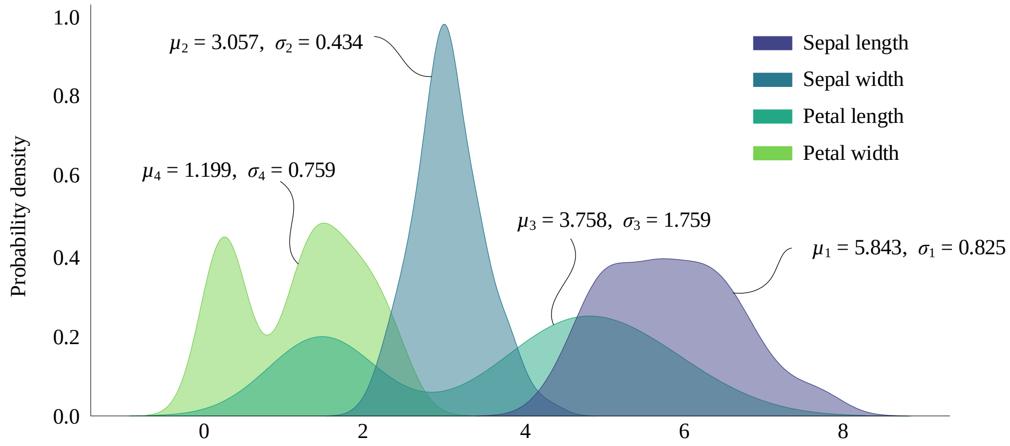


图 11. 鸢尾花数据四个特征上分布

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 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 协方差 $\operatorname{cov}(\boldsymbol{x}, \boldsymbol{y})$ 可以通过下式获得：

$$
\begin{aligned}
\operatorname{cov}(\boldsymbol{x},\boldsymbol{y}) &= \frac{\left(\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\boldsymbol{1}\right)^{\mathrm{T}}\left(\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})\boldsymbol{1}\right)}{n} = \frac{n\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{1}}{n^2} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\mathrm{E}(X)\right)\left(y_i-\mathrm{E}(Y)\right)}{n} = \frac{n\left(\sum_{i=1}^{n}x_iy_i\right)-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n^2}
\end{aligned}
\tag{39}
$$

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 $\boldsymbol{X}$，列向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的协方差有几种不同表达方式：

$$\operatorname{cov}\left(X_i, X_j\right)=\operatorname{cov}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)=\rho_{i,j}\sigma_i\sigma_j=\sigma_{i,j} \tag{40}$$

中心化矩阵

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

$$\sum_{i=1}^{n}\left(x_i-\mathrm{E}(X)\right)\left(y_i-\mathrm{E}(Y)\right)=(\boldsymbol{M}\boldsymbol{x})^{\mathrm{T}}\boldsymbol{M}\boldsymbol{y}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{y}=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{y} \tag{41}$$

联合 (35) 和 (41)，下式成立：

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

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：

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

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 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 相关性系数 $\operatorname{corr}(\boldsymbol{x},\boldsymbol{y})$ 计算式如下：

$$\operatorname{corr}(\boldsymbol{x},\boldsymbol{y})=\frac{\left(\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\right)\cdot\left(\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})\right)}{\left\|\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\right\|\left\|\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})\right\|}=\left(\frac{\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})}{\left\|\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\right\|}\right)\cdot\left(\frac{\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})}{\left\|\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})\right\|}\right) \tag{44}$$

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：

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

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：

$$\begin{aligned}\boldsymbol{x}\cdot\boldsymbol{y}&=\boldsymbol{y}\cdot\boldsymbol{x}\\ \operatorname{cov}(X,Y)&=\operatorname{cov}(Y,X)\end{aligned}\tag{46}$$

向量的模类似标准差：

$$\begin{aligned}\|\boldsymbol{x}\|&=\sqrt{\boldsymbol{x}\cdot\boldsymbol{x}}\\ \sigma_X&=\sqrt{\operatorname{var}(X)}=\sqrt{\operatorname{cov}(X,X)}\end{aligned}\tag{47}$$

向量之间夹角余弦值类似线性相关性系数：

$$\begin{aligned}\cos\theta&=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}\\ \rho_{X,Y}&=\operatorname{corr}(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\mathrm{E}\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X\sigma_Y}\end{aligned}\tag{48}$$

(48) 可以分别整理成如下等式：

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

此外，余弦定理可以用在向量内积和协方差上：

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

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 θ 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。

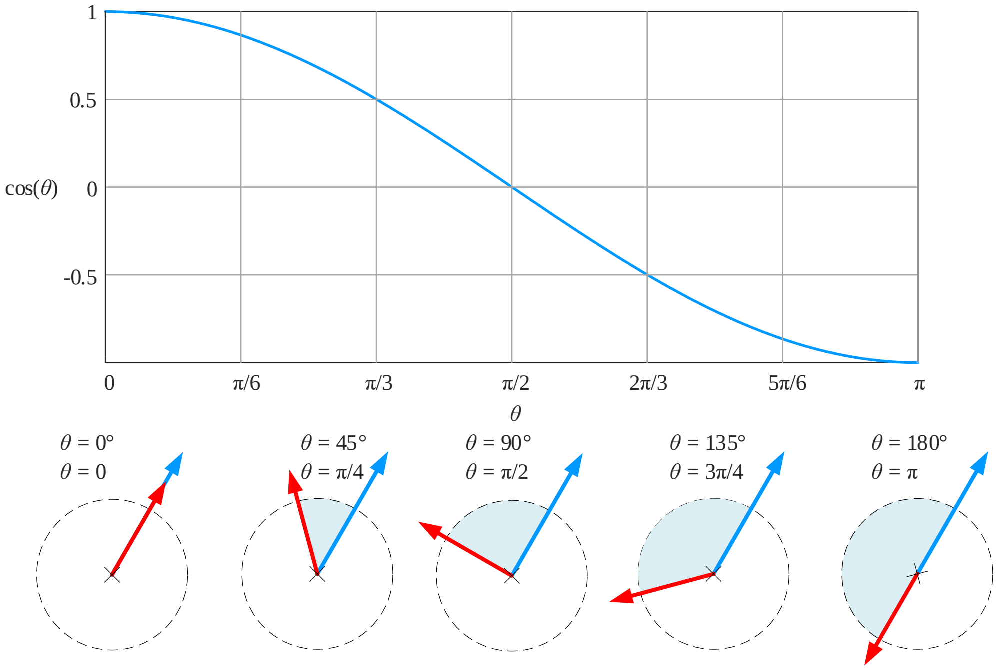


图 12. 余弦相似度

22.8 协方差矩阵和相关性系数矩阵

协方差矩阵

对于矩阵 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 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)：

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

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 $\boldsymbol{\Sigma}$ 对角线元素均为方差，其余元素为协方差。

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

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

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 $\boldsymbol{X}_c$ 代替 $\boldsymbol{X}-\mathrm{E}(\boldsymbol{X})$，(52) 可以写成：

$$\boldsymbol{\Sigma}=\frac{\overbrace{\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c}^{\text{Gram matrix}}}{n-1} \tag{53}$$

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 $\boldsymbol{X}_c$ 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：

$$\boldsymbol{\Sigma}=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{54}$$

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 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}$。将 (54) 代入 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}$，得到：

$$\begin{aligned}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}&=\boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}=\big(\underbrace{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}}_{\boldsymbol{y}}\big)^{\mathrm{T}}\boldsymbol{\Lambda}\big(\underbrace{\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}}_{\boldsymbol{y}}\big)=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}\\&=\lambda_1y_1^2+\lambda_2y_2^2+\cdots+\lambda_Dy_D^2=\sum_{j=1}^{D}\lambda_jy_j^2\end{aligned} \tag{55}$$

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

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

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}=\begin{bmatrix}x_1 & x_2\end{bmatrix}\begin{bmatrix}\sigma_{1,1} & \sigma_{1,2}\\ \sigma_{2,1} & \sigma_{2,2}\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}=\sigma_{1,1}x_1^2+\left(\sigma_{1,2}+\sigma_{2,1}\right)x_1x_2+\sigma_{2,2}x_2^2 \tag{56}$$

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

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}=\begin{bmatrix}y_1 & y_2\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}=\lambda_1y_1^2+\lambda_2y_2^2 \tag{57}$$

如图 13 所示，正是 (54) 中的 $\boldsymbol{V}$ 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。

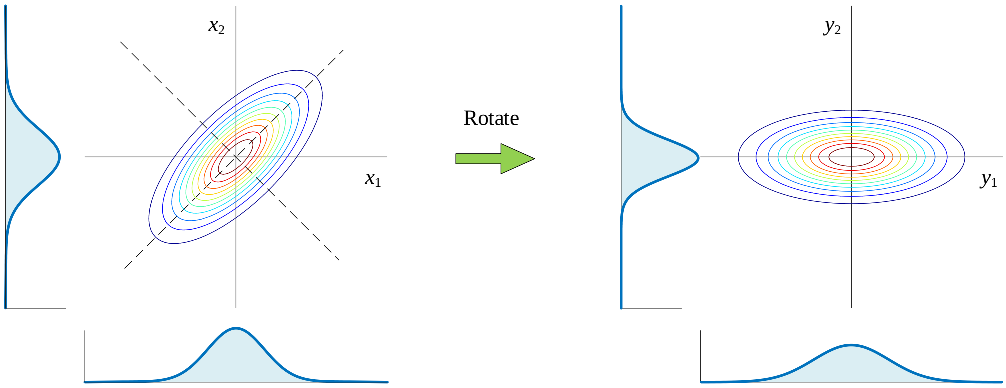


图 13. 旋转椭圆到正椭圆

相关性系数矩阵

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

$$\boldsymbol{P} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix} \tag{58}$$

$\boldsymbol{P}$ 和 $\boldsymbol{\Sigma}$ 的关系为：

$$\boldsymbol{\Sigma} = \boldsymbol{SPS} \tag{59}$$

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

再进一步，(58) 可以写成：

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

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 $\boldsymbol{Z}_X$。因此，(60) 可以写成：

$$\boldsymbol{P} = \frac{\boldsymbol{Z}_X^{\mathrm{T}} \boldsymbol{Z}_X}{n-1} \tag{61}$$

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

鸢尾花数据集

对于鸢尾花数据，它的协方差矩阵 $\boldsymbol{\Sigma}$ 为：

$$\boldsymbol{\Sigma} = \begin{array}{cc} \left[\begin{array}{cccc} 0.686 & -0.042 & 1.274 & 0.516 \\ -0.042 & 0.190 & -0.330 & -0.122 \\ 1.274 & -0.330 & 3.116 & 1.296 \\ 0.516 & -0.122 & 1.296 & 0.581 \end{array}\right] & \begin{array}{l} \leftarrow \text{Sepal length} \\ \leftarrow \text{Sepal width} \\ \leftarrow \text{Petal length} \\ \leftarrow \text{Petal width} \end{array} \\ \begin{array}{cccc} \text{Sepal length} & \text{Sepal width} & \text{Petal length} & \text{Petal width} \end{array} & \end{array} \tag{62}$$

鸢尾花数据的相关性系数矩阵 $\boldsymbol{P}$ 为：

$$\boldsymbol{P} = \begin{array}{cc} \left[\begin{array}{cccc} 1.000 & -0.118 & 0.872 & 0.818 \\ -0.118 & 1.000 & -0.428 & -0.366 \\ 0.872 & -0.428 & 1.000 & 0.963 \\ 0.818 & -0.366 & 0.963 & 1.000 \end{array}\right] & \begin{array}{l} \leftarrow \text{Sepal length} \\ \leftarrow \text{Sepal width} \\ \leftarrow \text{Petal length} \\ \leftarrow \text{Petal width} \end{array} \\ \begin{array}{cccc} \text{Sepal length} & \text{Sepal width} & \text{Petal length} & \text{Petal width} \end{array} & \end{array} \tag{63}$$

图 14 所示为 $\boldsymbol{\Sigma}$ 和 $\boldsymbol{P}$ 的热图。观察相关性系数矩阵 $\boldsymbol{P}$，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。

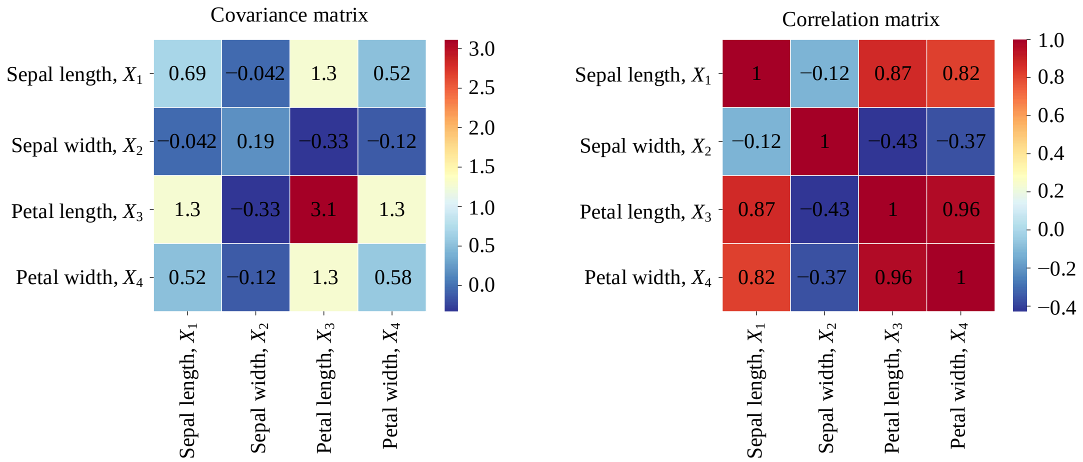


图 14. 协方差矩阵和相关性系数矩阵热图

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。

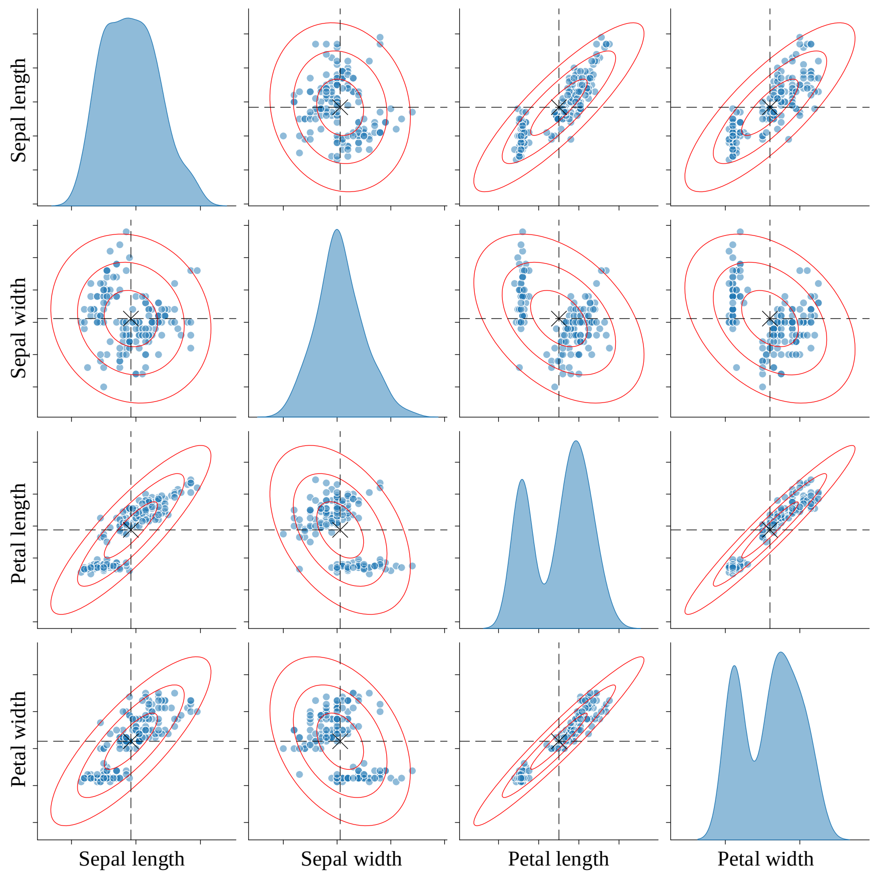


图 15. 协方差矩阵和椭圆的关系

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 $\boldsymbol{\Sigma}_k$ (如图 16) 和相关性系数矩阵 $\boldsymbol{P}_k$ (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。

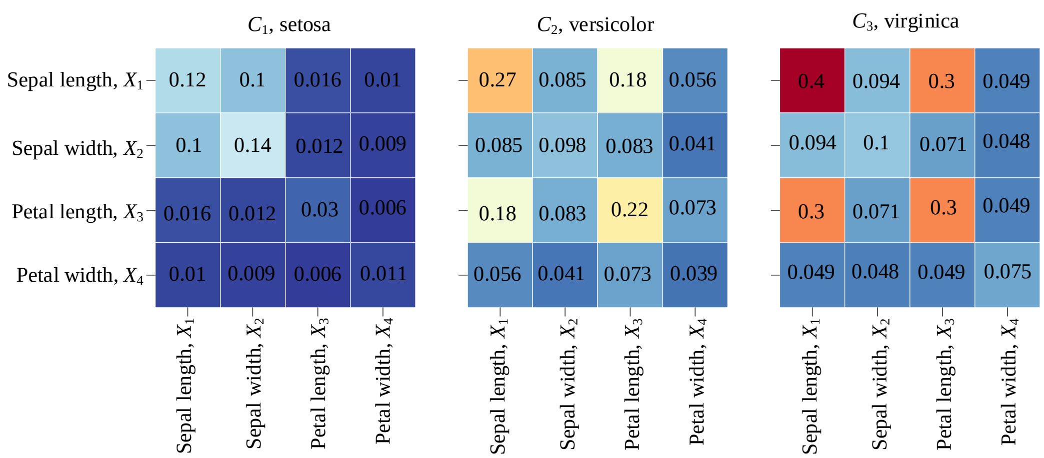


图 16. 协方差矩阵热图，考虑分类

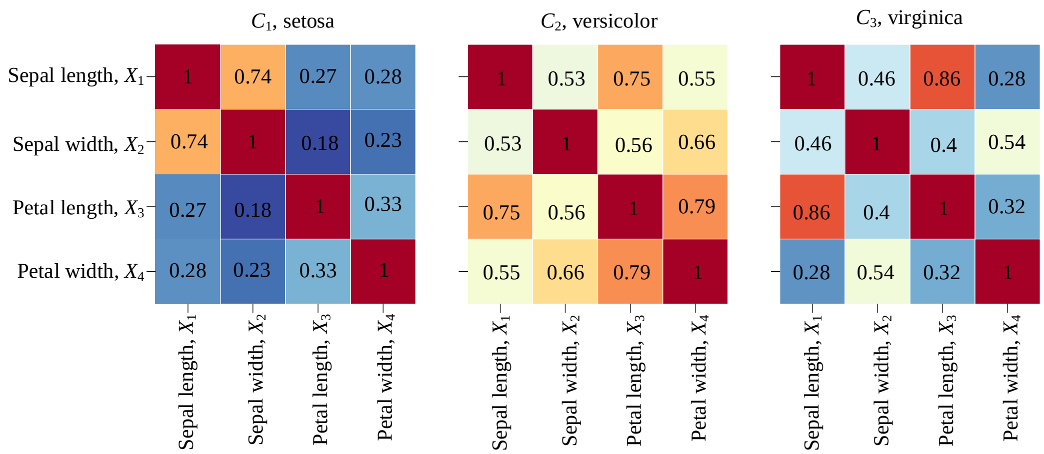


图 17. 相关性系数矩阵热图，考虑分类

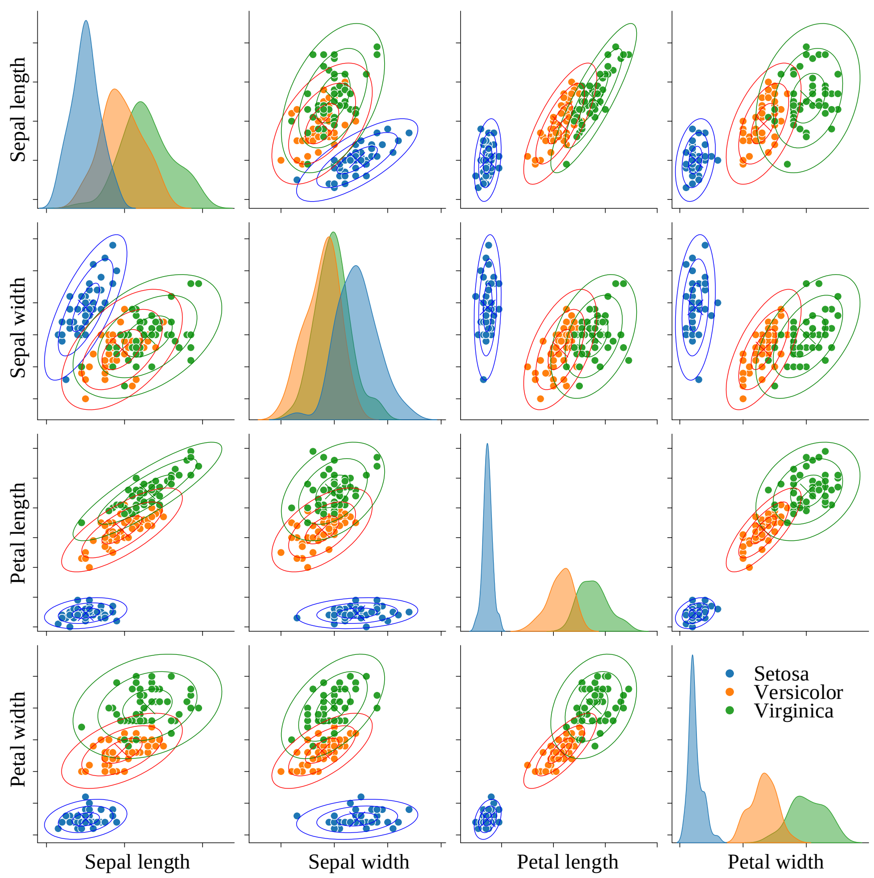


图 18. 协方差矩阵和椭圆的关系，考虑分类

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。

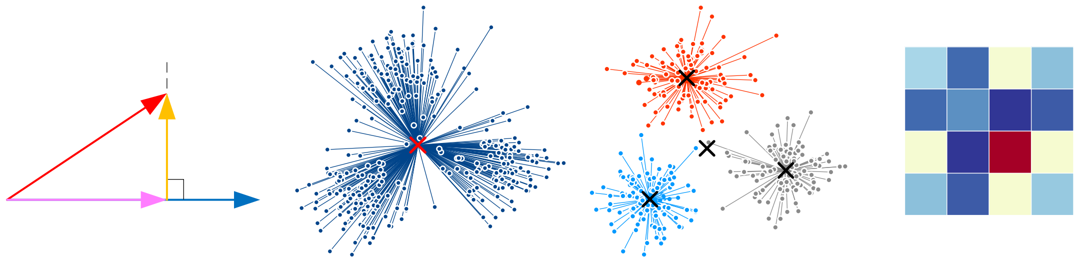


图 19. 总结本章重要内容的四幅图

本章介绍了两种和原始数据 $\boldsymbol{X}$ 形状相同的数据矩阵——中心化数据矩阵 $\boldsymbol{X}_c$、标准化数据矩阵 $\boldsymbol{Z}_{\boldsymbol{X}}$。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 Statical 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》。