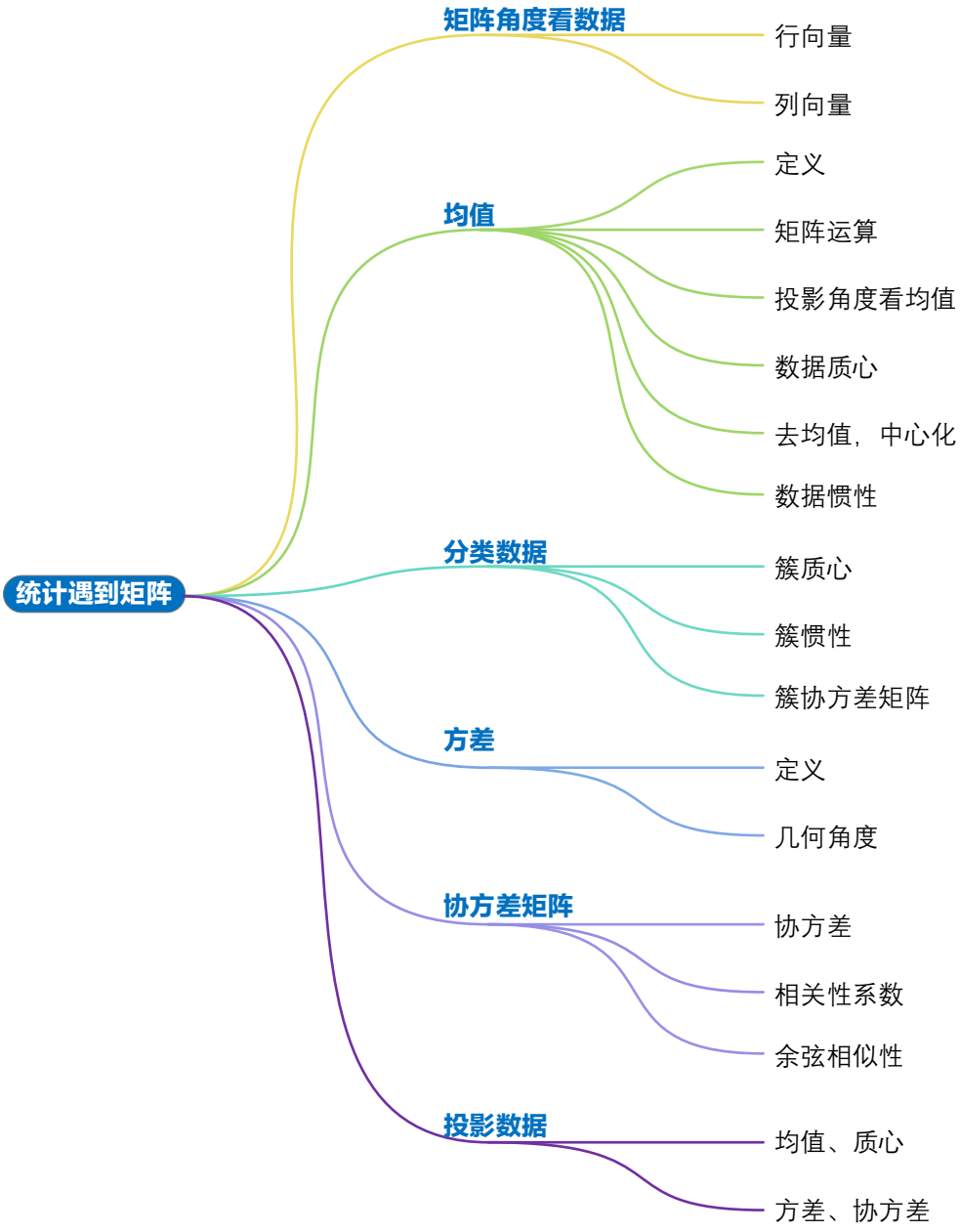
22 Statistics Meet Linear Algebra

数据与统计

从线性代数运算视角看统计

> 毫无争议的是，人类无法毫无错误地判断事物的真伪，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 向量或矩阵
- `seaborn.heatmap()` 绘制热图
- `seaborn.kdeplot()` 绘制核密度估计曲线



22.1 统计视角回看鸢尾花数据

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、数据投影结果的统计特征等内容。

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。它们分别对应 $\boldsymbol{X}$ 的四列。图 1 所示为使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。

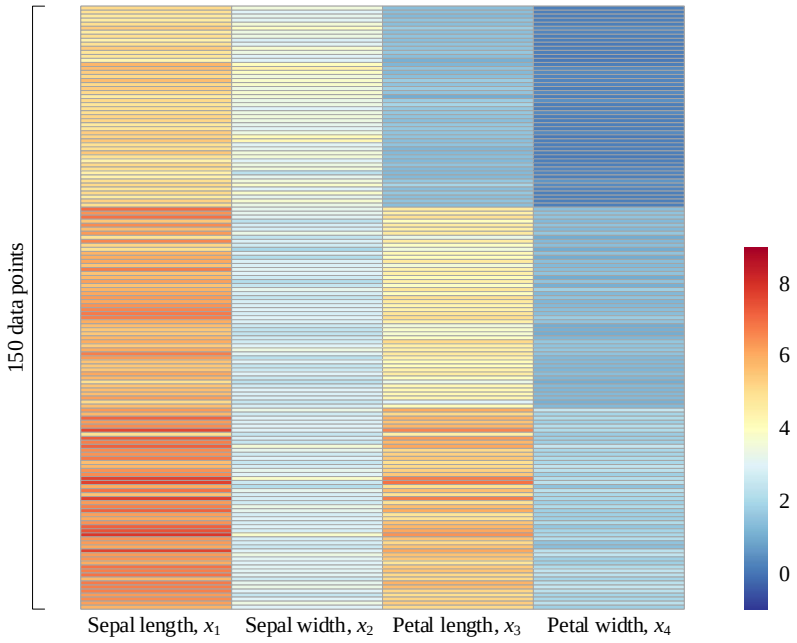


图 1. 鸢尾花数据，原始数据矩阵 $\boldsymbol{X}$

Bk4_Ch22_01.py 中 Bk4_Ch22_01_A 部分绘制图 1。

```
# Bk4_Ch22_01_A

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from sklearn.datasets import load_iris
```

```
# Load the iris data
iris_sns = sns.load_dataset("iris")
# A copy from Seaborn
iris = load_iris()
# A copy from Sklearn

X = iris.data
y = iris.target

feature_names = ['Sepal length, x1','Sepal width, x2',
                 'Petal length, x3','Petal width, x4']

# Convert X array to dataframe
X_df = pd.DataFrame(X, columns=feature_names)

#%% Heatmap of X

plt.close('all')
sns.set_style("ticks")

X = X_df.to_numpy();

# Visualize the heatmap of X

fig, ax = plt.subplots()
ax = sns.heatmap(X,
                 cmap='RdYlBu_r',
                 xticklabels=list(X_df.columns),
                 cbar_kws={"orientation": "vertical"},
                 vmin=-1, vmax=9)
plt.title('X')
```

22.2 均值的矩阵运算

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：

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

回顾求解随机变量 X_j 的期望值 (均值) $\mathrm{E}(X_j)$ 运算：

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

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

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

其中，$\boldsymbol{1}$ 为全 1 列向量，行数和 $\boldsymbol{x}_j$ 一致；E() 计算期望值/均值。

(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。

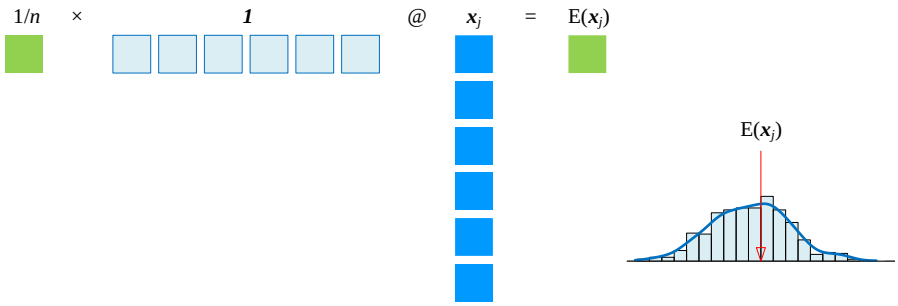


图 2. 计算 $\boldsymbol{x}_j$期望值/均值

我们可以利用矩阵运算分别得到鸢尾花的四个特征的期望值：

$$\begin{cases} \mathrm{E}\left(\boldsymbol{x}_1\right) = \mu_1 = \underset{\text{Sepal length, } x_1}{5.843} \\ \mathrm{E}\left(\boldsymbol{x}_2\right) = \mu_2 = \underset{\text{Sepal width, } x_2}{3.057} \\ \mathrm{E}\left(\boldsymbol{x}_3\right) = \mu_3 = \underset{\text{Petal length, } x_3}{3.758} \\ \mathrm{E}\left(\boldsymbol{x}_4\right) = \mu_4 = \underset{\text{Petal width, } x_4}{1.199} \end{cases} \tag{5}$$

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：

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

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，没有解释的向量分量为：

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

这部分垂直于 $\boldsymbol{1}$，也就是：

$$\mathbf{1}^{\mathrm{T}}\left(\boldsymbol{x}_j-\mathrm{proj}_{\mathbf{1}}\left(\boldsymbol{x}_j\right)\right)=\mathbf{1}^{\mathrm{T}}\left(\boldsymbol{x}_j-\frac{\boldsymbol{x}_j^{\mathrm{T}}\mathbf{1}}{n}\mathbf{1}\right)=\mathbf{1}^{\mathrm{T}}\boldsymbol{x}_j-\frac{\boldsymbol{x}_j^{\mathrm{T}}\mathbf{1}}{n}\mathbf{1}^{\mathrm{T}}\mathbf{1}_n=\mathbf{1}^{\mathrm{T}}\boldsymbol{x}_j-\boldsymbol{x}_j^{\mathrm{T}}\mathbf{1}=0 \tag{8}$$

也就是说，均值作为一个统计量，它能解释 $\boldsymbol{x}_j$ 一部分特征，但并不能解释所有特征。

均值没有解释的这部分特征，$\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_j)\mathbf{1}$，将在标准差 (方差平方根) 中加以解释。

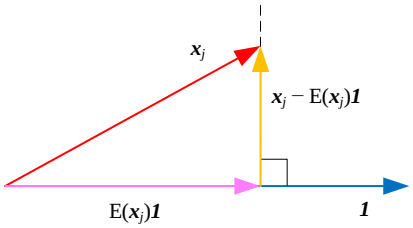


图 3. 投影角度看期望值

两个极端例子

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

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

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

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

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

也就是意味着 $\mathrm{E}(\boldsymbol{x}_j)=0$。

对于最小二乘法线性回归，$\boldsymbol{x}_j-\mathrm{E}(\boldsymbol{x}_j)\mathbf{1}$ 垂直于 $\mathbf{1}$ 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 $\boldsymbol{X}$ 的每一特征均值构成的向量，我们定义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。

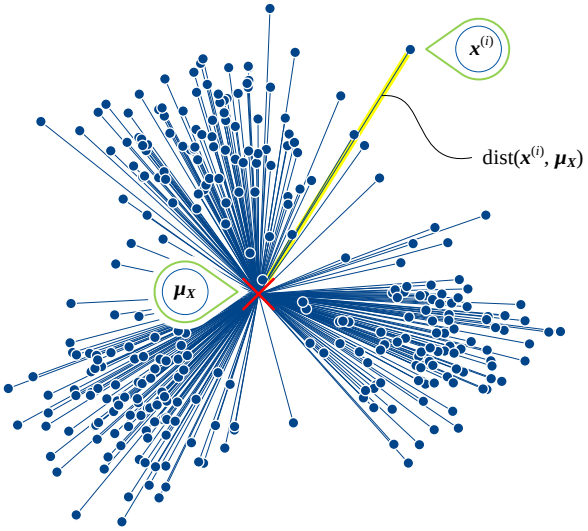


图 4. 平面上数据矩阵 $\boldsymbol{X}$ 质心位置

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。

在多元高斯分布中，我们会用列向量 $\boldsymbol{\mu}_X$：

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

上式中，几何角度来看，$\boldsymbol{x}-\boldsymbol{\mu}_X$ 相当于“平移”。

前文介绍过，$\boldsymbol{\mu}_X$ 可以通过如下矩阵运算获得：

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

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。

举个例子，鸢尾花数据质心位置：

$$\boldsymbol{\mu}_X = \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \tag{14}$$

整理 (13) 上式得到两个等式：

$$\begin{cases} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{1} = n\boldsymbol{\mu}_X \\ \boldsymbol{1}^{\mathrm{T}}\boldsymbol{X} = n\left(\boldsymbol{\mu}_X\right)^{\mathrm{T}} \end{cases} \tag{15}$$

我们将会在一些运算中用到这两个等式。

行向量

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

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

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。

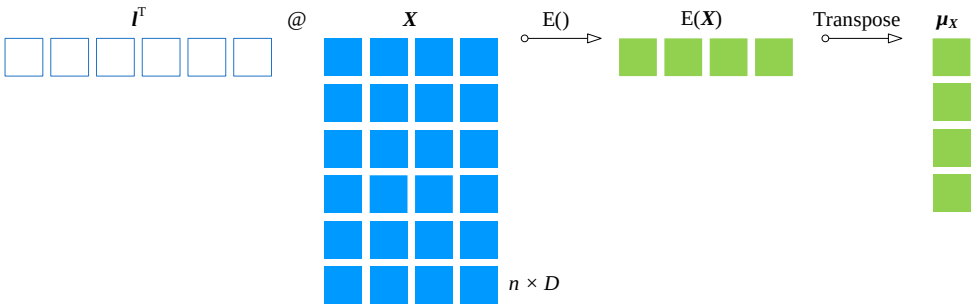


图 5. 计算 $\boldsymbol{X}$ 样本数据的质心 $\boldsymbol{\mu}_X$

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：

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

上式 $\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}$ 为全 1 列向量和其转置乘积；$\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\right)$ 为中心化矩阵。

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，对应矩阵运算为：

$$\boldsymbol{X}_c = \boldsymbol{X} - \boldsymbol{1}\left(\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{1}}{n}\right)^{\mathrm{T}} = \boldsymbol{X} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{X} = \left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\right)\boldsymbol{X} \tag{19}$$

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

为了方便表达，我们可以利用广播原则来中心化 $\boldsymbol{X}$，即用原始数据矩阵 $\boldsymbol{X}$ 减去行向量 $\mathrm{E}(\boldsymbol{X})$：

$$\boldsymbol{X}_c = \boldsymbol{X} - \mathrm{E}(\boldsymbol{X}) \tag{20}$$

图 6 所示为鸢尾花数据去均值后热图。中心化后，数据 $\boldsymbol{X}_c$ 质心位于原点。

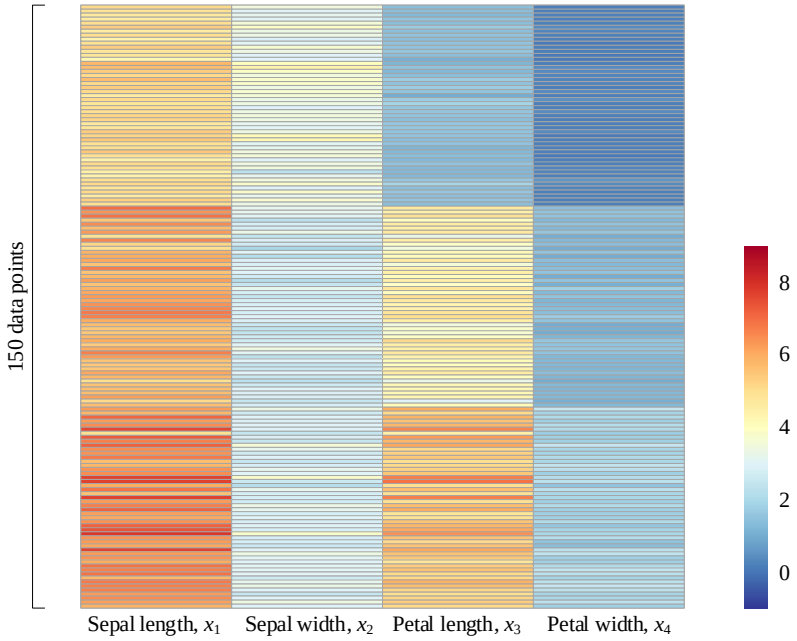


图 6. 鸢尾花数据矩阵 $\boldsymbol{X}$，去均值

标准化：平移 + 缩放

在中心化的基础上，我们可以进一步对数据进行标准化 (standardization 或 z-score normalization)。计算过程为，对原始数据先去均值，然后再除以标准差：

$$\boldsymbol{Z}_{\boldsymbol{X}} = \boldsymbol{X}_c \boldsymbol{S}^{-1} = \left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)\boldsymbol{S}^{-1} \tag{21}$$

其中，缩放矩阵 $\boldsymbol{S}$ 为：

$$\boldsymbol{S} = \mathrm{diag}\left(\mathrm{diag}\left(\boldsymbol{\Sigma}\right)\right)^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix} \tag{22}$$

处理得到的数值实际上是原始数据的 z 分数 (z score)，表达若干倍的标准差偏移。

比如说，处理得到的数值为 3，也就是说这个数据是距离均值 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、比较单位不同的特征样本数据。

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。

惯性

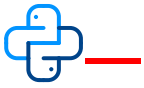
数据**惯性** (inertia) 可以用来描述样本数据紧密程度，惯性实际上就是**总离差平方和** (Sum of Squares for Deviations, SSD)，定义如下：

$$\mathrm{SSD}\left(\boldsymbol{X}\right) = \sum_{i=1}^{n} \left\| \boldsymbol{x}^{(i)} - \mathrm{E}\left(\boldsymbol{X}\right) \right\|_2^2 = \sum_{i=1}^{n} \mathrm{dist}\left(\boldsymbol{x}^{(i)}, \mathrm{E}\left(\boldsymbol{X}\right)\right)^2 \tag{23}$$

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

(23) 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：

$$\mathrm{SSD}\left(\boldsymbol{X}\right) = \mathrm{trace}\left(\boldsymbol{X}_c^{\mathrm{T}} \boldsymbol{X}_c\right) = \mathrm{trace}\left(\left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)^{\mathrm{T}} \left(\boldsymbol{X} - \mathrm{E}\left(\boldsymbol{X}\right)\right)\right) \tag{24}$$



Bk4_Ch22_01.py 中 Bk4_Ch22_01_B 部分绘制图 6 并计算 SSD。

```
# Bk4_Ch22_01_B

#%% centroid of data matrix, X
v_1 = np.ones((len(X),1))

E_X = v_1.T@X/len(X)
# validate: X.mean(axis = 0)
```

```
#%% Demean, centralize

X_demean = X_df.sub(X_df.mean())


fig, ax = plt.subplots()
ax = sns.heatmap(X_demean,
                 cmap='RdYlBu_r',
                 xticklabels=list(X_df.columns),
                 cbar_kws={"orientation": "vertical"},
                 vmin=-3, vmax=3)
plt.title('$X_{demean}$')

#%% SSD

SSD = (np.linalg.norm(X - E_X, axis = 1)**2).sum()
# validate: ((X - E_X)**2).sum()
# use trace: np.trace((X - E_X).T@(X - E_X))
```

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。

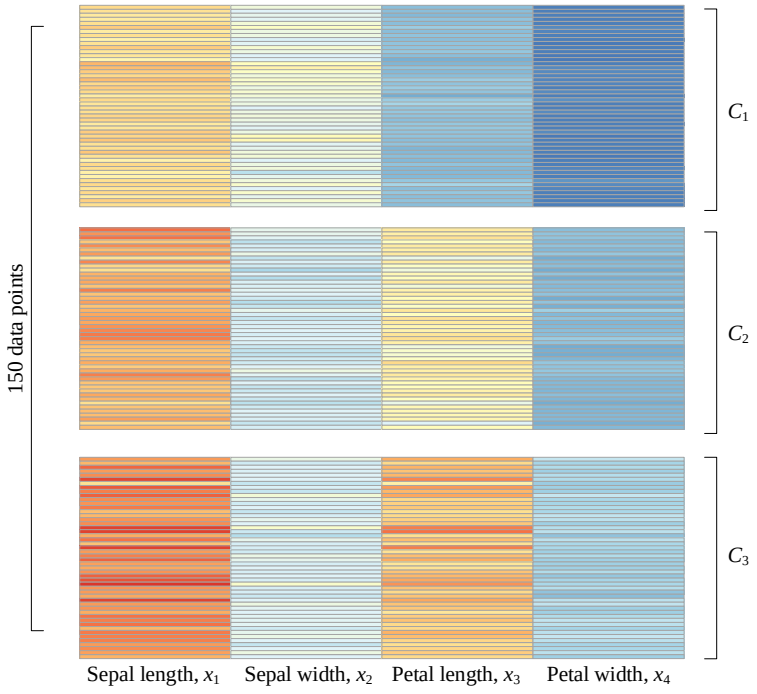


图 7. 鸢尾花数据分为三类

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：

$$\boldsymbol{\mu}_k = \frac{1}{\text{count}\left(C_k\right)} \sum_{i \in C_k} \boldsymbol{x}^{(i)} \tag{25}$$

公式看上去很复杂，实际道理其实很简单。

简单翻译一下，对于属于某个标签 C_k 的所有样本数据 $\boldsymbol{x}^{(i)}$ ($i \in C_k$)，求解其各个特征位置平均值，构造成一个新的列向量 $\boldsymbol{\mu}_k$。图 8 所示为样本数据质心 $\boldsymbol{\mu}_X$，和三类数据各自的质心 $\boldsymbol{\mu}_1$、$\boldsymbol{\mu}_2$ 和 $\boldsymbol{\mu}_3$ 之间的关系。

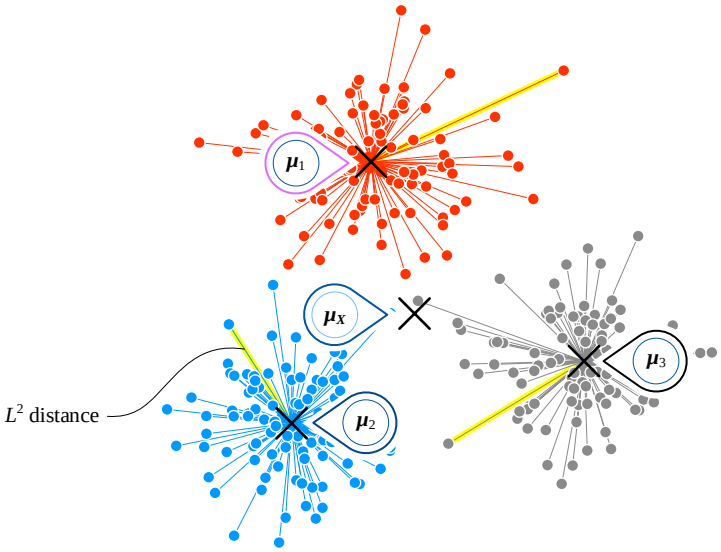


图 8. 样本数据质心 $\boldsymbol{\mu}_X$，和三类数据各自的质心 $\boldsymbol{\mu}_1$、$\boldsymbol{\mu}_2$ 和 $\boldsymbol{\mu}_3$

举个简单例子。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：

$$\boldsymbol{x}^{(2)} = \begin{bmatrix} 2 & 3 \end{bmatrix}^{\mathrm{T}}, \quad \boldsymbol{x}^{(5)} = \begin{bmatrix} 3 & 1 \end{bmatrix}^{\mathrm{T}}, \quad \boldsymbol{x}^{(6)} = \begin{bmatrix} -2 & 2 \end{bmatrix}^{\mathrm{T}}, \quad \boldsymbol{x}^{(9)} = \begin{bmatrix} 1 & 6 \end{bmatrix}^{\mathrm{T}} \tag{26}$$

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

$$\begin{aligned} \boldsymbol{\mu}_{C1} &= \frac{1}{\text{count}\left(C_1\right)} \sum_{i \in C_1} \boldsymbol{x}^{(i)} = \frac{1}{\text{count}\left(C_1\right)} \left(\boldsymbol{x}^{(2)} + \boldsymbol{x}^{(5)} + \boldsymbol{x}^{(6)} + \boldsymbol{x}^{(9)} \right) \\ &= \frac{1}{4} \left(\begin{bmatrix} 2 & 3 \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} 3 & 1 \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} -2 & 2 \end{bmatrix}^{\mathrm{T}} + \begin{bmatrix} 1 & 6 \end{bmatrix}^{\mathrm{T}} \right) \\ &= \begin{bmatrix} 1 & 3 \end{bmatrix}^{\mathrm{T}} \end{aligned} \tag{27}$$

以鸢尾花数据为例，计算簇质心就是对图 7 三组数据分别计算质心。图 9 不同颜色的 × 代表鸢尾花的簇质心位置。

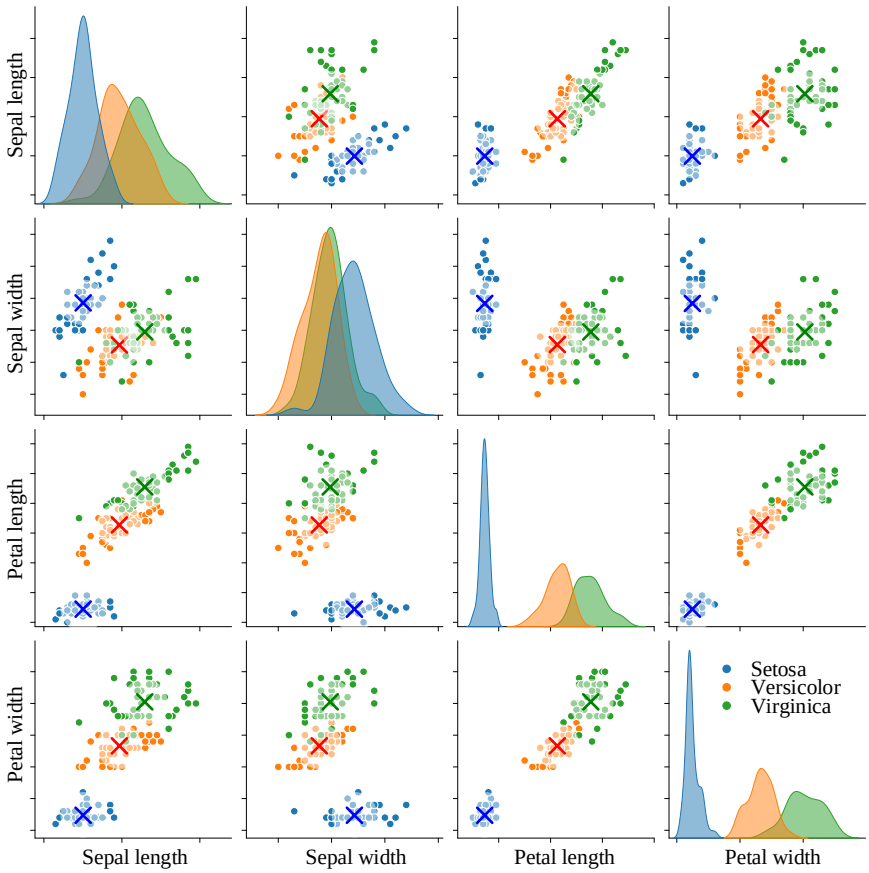


图 9. 鸢尾花数据簇质心位置

簇惯量

簇惯量 (cluster inertia) 定义如下：

$$\mathrm{SSD}(C_k) = \sum_{i \in C_k} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{\mu}_k \right\|_2^2 \tag{28}$$

(28) 相当于任意一簇 C_k 样本点距离簇质心 $\boldsymbol{\mu}_k$ 欧氏距离平方和。

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：

$$\mathrm{var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}{n^2} \tag{29}$$

对于样本来说，随机变量 X 方差的计算式为：

$$\mathrm{var}(X)=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{30}$$

如果不考虑总体和样本在计算方差上的差异，var(X) 和 E(X^2)及 E(X)的关系如下：

$$\begin{aligned}\mathrm{var}(X)&=\mathrm{E}\left((X-\mathrm{E}(X))^2\right)\\&=\mathrm{E}\left(X^2-2X\,\mathrm{E}(X)+\mathrm{E}(X)^2\right)\\&=\mathrm{E}\left(X^2\right)-\mathrm{E}(X)^2\end{aligned} \tag{31}$$

$\boldsymbol{x}$ 为随机变量 X 对应的列向量，利用下式计算获得 var($\boldsymbol{x}$)：

$$\mathrm{var}(\boldsymbol{x})=\mathrm{E}(\boldsymbol{x}\odot\boldsymbol{x})-\mathrm{E}(\boldsymbol{x})^2 \tag{32}$$

其中，⊙为向量逐项积运算符。

进一步展开 (32) 得到：

$$\begin{aligned}\mathrm{var}(\boldsymbol{x})&=\mathrm{E}(\boldsymbol{x}\odot\boldsymbol{x})-\mathrm{E}(\boldsymbol{x})^2\\&=\frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}{n}-\left(\frac{1}{n}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}\right)^2=\frac{n\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{1}}{n^2}\\&=\frac{n\left(\sum_{i=1}^{n}x_i^2\right)-\left(\sum_{i=1}^{n}x_i\right)^2}{n^2}\end{aligned} \tag{33}$$

根据 (32)，还可以得到如下等式：

$$\mathrm{E}(\boldsymbol{x}\odot\boldsymbol{x})=\mathrm{var}(\boldsymbol{x})+\mathrm{E}(\boldsymbol{x})^2 \tag{34}$$

观察 (33)，我们还可以得到得到 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}$、均值、方差三者关系。

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}=n\left(\mathrm{var}(\boldsymbol{x})+\mathrm{E}(\boldsymbol{x})^2\right) \tag{35}$$

向量视角

前文介绍了均值的投影视角，将图 10 视角应用到本节标准差。

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

E($\boldsymbol{x}$)$\boldsymbol{1}$ 和 $\boldsymbol{1}$ 平行，而 $\boldsymbol{x}$ – E($\boldsymbol{x}$)$\boldsymbol{1}$ 和 $\boldsymbol{1}$ 垂直。而向量 $\boldsymbol{x}$ – E($\boldsymbol{x}$)$\boldsymbol{1}$ 的模的平方就是 n 倍方差。

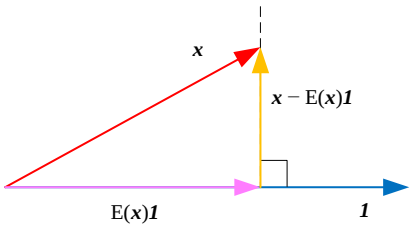


图 10. 投影角度看方差和标准差

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：

$$\operatorname{var}\left(X_j\right)=\operatorname{var}\left(\boldsymbol{x}_j\right)=\sigma_j^2=\sigma_{j,j} \tag{36}$$

计算鸢尾花数据 $\boldsymbol{X}$ 每一列标准差，以行向量表达：

$$\boldsymbol{\sigma}_{\boldsymbol{X}}=\begin{bmatrix} 0.825 & 0.434 & 1.759 & 0.759 \\ \text{Sepal length, } x_1 & \text{Sepal width, } x_2 & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{bmatrix} \tag{37}$$

$\boldsymbol{X}$ 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 的含义是核密度估计 (Kernel Density Estimation, KDE)，这是本系列丛书《概率统计》一册要讲解的话题。

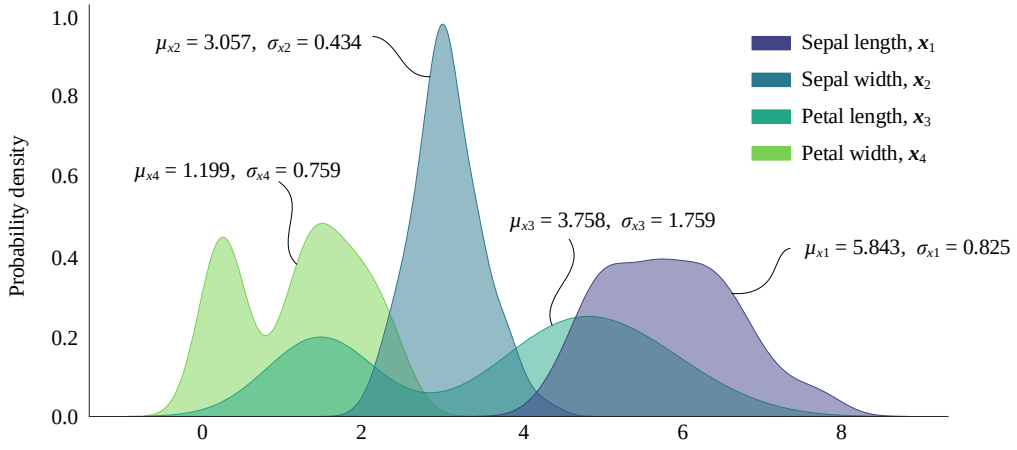
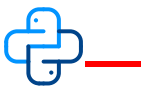


图 11. 鸢尾花数据四个特征上分布



Bk4_Ch22_01.py 中 Bk4_Ch22_01_C 部分绘制图 11。

```
# Bk4_Ch22_01_C

# distribution of column features of X

fig, ax = plt.subplots()
```

```
sns.kdeplot(data=X_demean,fill=True,
            common_norm=False,
            alpha=.3, linewidth=1,
            palette = "viridis")
plt.title('Distribution of $X_{demean}$ columns')
```

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 $\text{cov}(\boldsymbol{x},\boldsymbol{y})$ 可以通过下式获得：

$$\begin{aligned}\text{cov}(\boldsymbol{x},\boldsymbol{y}) &= \text{E}\left[(\boldsymbol{x}-\text{E}(\boldsymbol{x}))(\boldsymbol{y}-\text{E}(\boldsymbol{y}))\right] \\ &= \text{E}\left[\boldsymbol{x}\odot\boldsymbol{y}-\boldsymbol{x}\text{E}(\boldsymbol{y})-\text{E}(\boldsymbol{x})\boldsymbol{y}+\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y})\right] \\ &= \text{E}(\boldsymbol{x}\odot\boldsymbol{y})-\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y})-\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y})+\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y}) \\ &= \text{E}(\boldsymbol{x}\odot\boldsymbol{y})-\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y})\end{aligned} \tag{38}$$

上式中 $\text{E}(\boldsymbol{x}\odot\boldsymbol{y})$ 是 $\boldsymbol{x}$ 向量和 $\boldsymbol{y}$ 向量对应元素相乘之后再求平均数。

整理 (38)，可以得到如下等式：

$$\text{E}(\boldsymbol{x}\odot\boldsymbol{y}) = \text{cov}(\boldsymbol{x},\boldsymbol{y})+\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y}) \tag{39}$$

用向量内积方式来写，协方差为：

$$\text{cov}(\boldsymbol{x},\boldsymbol{y}) = \frac{(\boldsymbol{x}-\text{E}(\boldsymbol{x}))\cdot(\boldsymbol{y}-\text{E}(\boldsymbol{y}))}{n} \tag{40}$$

将具体数值代入 (38)，并整理得到：

$$\begin{aligned}\text{cov}(\boldsymbol{x},\boldsymbol{y}) &= \text{E}(\boldsymbol{x}\odot\boldsymbol{y})-\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y}) \\ &= \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}}{n}-\frac{1}{n}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{l}\frac{1}{n}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{l} = \frac{n\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{l}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{l}}{n^2} \\ &= \frac{n\left(\sum_{i=1}^{n}x_iy_i\right)-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n^2}\end{aligned} \tag{41}$$

观察上式得到 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ 和 $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x}$ 解析式：

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = n\left(\text{cov}(\boldsymbol{x},\boldsymbol{y})+\text{E}(\boldsymbol{x})\text{E}(\boldsymbol{y})\right) \tag{42}$$

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

$$\text{cov}(X_i,X_j) = \text{cov}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \rho_{i,j}\sigma_i\sigma_j = \sigma_{i,j} \tag{43}$$

显然，$\boldsymbol{x}$ 和全 1 列向量 $\boldsymbol{1}$ 的协方差为 0：

$$\begin{aligned}\operatorname{cov}(\boldsymbol{x},\boldsymbol{1}) &= \mathrm{E}(\boldsymbol{x}\odot\boldsymbol{1}) - \mathrm{E}(\boldsymbol{x})\mathrm{E}(\boldsymbol{1}) \\ &= \mathrm{E}(\boldsymbol{x}) - \mathrm{E}(\boldsymbol{x}) \\ &= 0\end{aligned} \tag{44}$$

相关性系数

随机变量 X 和 Y 相关性系数的定义为：

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{45}$$

相关性系数可以看做是随机变量 z 分数的协方差。

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

$$\operatorname{corr}(\boldsymbol{x},\boldsymbol{y}) = \frac{(\boldsymbol{x}-\mathrm{E}(\boldsymbol{x}))\cdot(\boldsymbol{y}-\mathrm{E}(\boldsymbol{y}))}{\|\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\|\|\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})\|} = \left(\frac{\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})}{\|\boldsymbol{x}-\mathrm{E}(\boldsymbol{x})\|}\right)\cdot\left(\frac{\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})}{\|\boldsymbol{y}-\mathrm{E}(\boldsymbol{y})\|}\right) \tag{46}$$

相信大家已经在上式中，看到“平移”和“缩放”。也就是说，我们将 [$\boldsymbol{x}$, $\boldsymbol{y}$] 中心化，然后再对列向量单位化。

两个单位向量的内积结果就是向量夹角的余弦值，这边引出我们本节最后要讲的内容。

余弦相似性

余弦相似性 (cosine similarity) 通过两个向量的夹角的余弦值来度量它们的相似性：

$$\text{cosine similarity} = \cos(\theta) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \tag{47}$$

向量内积和协方差有诸多相似之处。向量内积和协方差都满足交换律：

$$\begin{aligned}\boldsymbol{x}\cdot\boldsymbol{y} &= \boldsymbol{y}\cdot\boldsymbol{x} \\ \operatorname{cov}(X,Y) &= \operatorname{cov}(Y,X)\end{aligned} \tag{48}$$

向量的模类似标准差：

$$\begin{aligned}\|\boldsymbol{x}\| &= \sqrt{\boldsymbol{x}\cdot\boldsymbol{x}} \\ \sigma_X &= \sqrt{\operatorname{var}(X)} = \sqrt{\operatorname{cov}(X,X)}\end{aligned} \tag{49}$$

向量之间夹角余弦值类似线性相关性系数：

$$\begin{aligned}\cos\theta &= \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} \\ \rho_{X,Y} &= \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\mathrm{E}\left((X-\mu_X)(Y-\mu_Y)\right)}{\sigma_X\sigma_Y}\end{aligned} \tag{50}$$

(50) 可以整理成如下等式：

$$\begin{aligned} \boldsymbol{x}\cdot\boldsymbol{y} &= \cos\theta\|\boldsymbol{x}\|\|\boldsymbol{y}\| \\ \operatorname{cov}(X,Y) &= \rho_{X,Y}\sigma_X\sigma_Y \end{aligned} \tag{51}$$

此外，余弦定理可以用在向量内积和协方差上：

$$\begin{aligned} &\|\boldsymbol{x}+\boldsymbol{y}\|^2 = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 + 2\|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta \\ &\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X,Y) \\ &\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho_{X,Y}\sigma_X\sigma_Y \\ &\operatorname{var}(aX+bY) = a^2\operatorname{var}(X) + b^2\operatorname{var}(Y) + 2ab\operatorname{cov}(X,Y) \end{aligned} \tag{52}$$

值得一提的是，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，方差和协方差运算中都存在平移。

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 所示为余弦相似性和夹角 θ 关系。

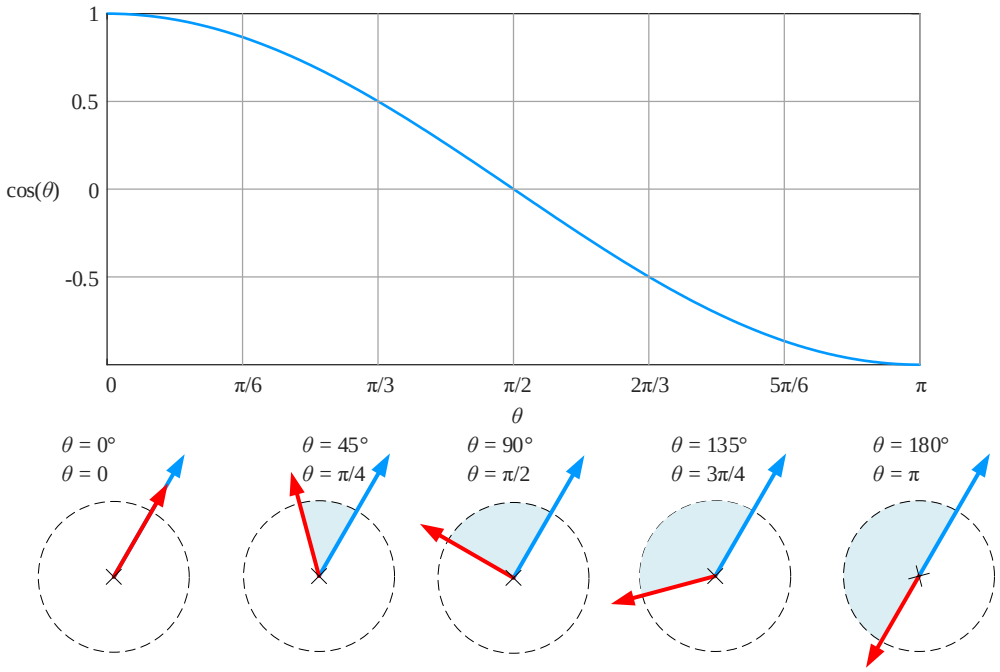


图 12. 余弦相似度

22.8 协方差矩阵和相关性系数矩阵

协方差矩阵

对于矩阵 $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$ 每两列之间的协方差可以构造得到**协方差矩阵** (covariance matrix)：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \operatorname{cov}(\boldsymbol{x}_1, \boldsymbol{x}_1) & \operatorname{cov}(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & \operatorname{cov}(\boldsymbol{x}_1, \boldsymbol{x}_D) \\ \operatorname{cov}(\boldsymbol{x}_2, \boldsymbol{x}_1) & \operatorname{cov}(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & \operatorname{cov}(\boldsymbol{x}_2, \boldsymbol{x}_D) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{cov}(\boldsymbol{x}_D, \boldsymbol{x}_1) & \operatorname{cov}(\boldsymbol{x}_D, \boldsymbol{x}_2) & \cdots & \operatorname{cov}(\boldsymbol{x}_D, \boldsymbol{x}_D) \end{bmatrix} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D,1} & \sigma_{D,2} & \cdots & \sigma_{D,D} \end{bmatrix} \tag{53}$$

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，这是因为上式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：

$$\boldsymbol{\Sigma} = \frac{\left(\underbrace{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})}_{\text{Centered}} \right)^{\mathrm{T}} \left(\underbrace{\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})}_{\text{Centered}} \right)}{n-1} \tag{54}$$

对于总体，上式的分母改为 n。

特征值分解

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前文介绍的二次型 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}$ 。

对 $\boldsymbol{\Sigma}$ 进行特征值分解，得到：

$$\boldsymbol{\Sigma} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} \tag{55}$$

将 (55) 代入 $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}$，得到：

$$\begin{aligned} \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x} &= \boldsymbol{x}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x} \\ &= \underbrace{\left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}\right)}_{\boldsymbol{y}}^{\mathrm{T}} \boldsymbol{\Lambda} \underbrace{\left(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{x}\right)}_{\boldsymbol{y}} \\ &= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y} \\ &= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_D y_D^2 = \sum_{j=1}^{D} \lambda_j y_j^2 \end{aligned} \tag{56}$$

大家是否眼前一亮，$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}$ 就是正椭球，这意味着 $\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{y}$ 为旋转椭球。特别地，当 $D = 2$ 时，$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x}$ 旋转椭圆：

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \sigma_{1,1}x_1^2 + \left(\sigma_{1,2} + \sigma_{2,1}\right)x_1x_2 + \sigma_{2,2}x_2^2 \tag{57}$$

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

$$\boldsymbol{y}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{y}=\begin{bmatrix}y_1 & y_2\end{bmatrix}\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}y_1\\ y_2\end{bmatrix}=\lambda_1 y_1^2+\lambda_2 y_2^2 \tag{58}$$

而正是 (55) 中的 $\boldsymbol{V}$ 完成正椭圆到旋转椭圆的“旋转”。

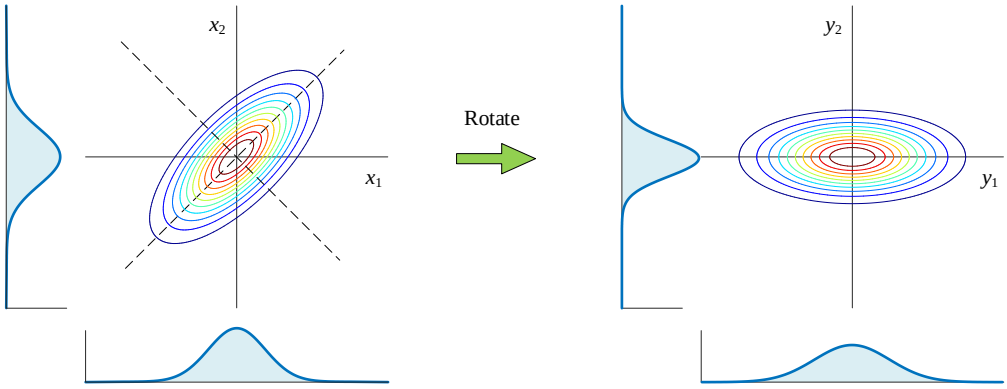


图 13. 旋转椭圆到正椭圆

相关性系数矩阵

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

$$\boldsymbol{P}=\begin{bmatrix}1 & \rho_{1,2} & \cdots & \rho_{1,D}\\ \rho_{1,2} & 1 & \cdots & \rho_{2,D}\\ \vdots & \vdots & \ddots & \vdots\\ \rho_{1,D} & \rho_{2,D} & \cdots & 1\end{bmatrix} \tag{59}$$

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

$$\boldsymbol{\Sigma}=\boldsymbol{SPS} \tag{60}$$

$\boldsymbol{S}$ 就是 (22) 定义的缩放矩阵，$\boldsymbol{S}$ 是个对角方阵。

鸢尾花数据集

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

$$\boldsymbol{\Sigma}=\begin{bmatrix}0.686 & -0.042 & 1.274 & 0.516\\ -0.042 & 0.190 & -0.330 & -0.122\\ 1.274 & -0.330 & 3.116 & 1.296\\ 0.516 & \underbrace{-0.122}_{\text{Sepal width, } x_2} & 1.296 & 0.581\\ \text{Sepal length, } x_1 & & \text{Petal length, } x_3 & \text{Petal width, } x_4\end{bmatrix}\begin{matrix}\leftarrow \text{Sepal length, } x_1\\ \leftarrow \text{Sepal width, } x_2\\ \leftarrow \text{Petal length, } x_3\\ \leftarrow \text{Petal width, } x_4\end{matrix} \tag{61}$$

鸢尾花数据的相关性系数矩阵 $\boldsymbol{P}$ 为：

$$
\boldsymbol{P} = \begin{bmatrix} 1.000 & -0.118 & 0.872 & 0.818 \\ -0.118 & 1.000 & -0.428 & -0.366 \\ 0.872 & -0.428 & 1.000 & 0.963 \\ 0.818 & -0.366 & 0.963 & 1.000 \end{bmatrix} \begin{matrix} \leftarrow \text{Sepal length, } x_1 \\ \leftarrow \text{Sepal width, } x_2 \\ \leftarrow \text{Petal length, } x_3 \\ \leftarrow \text{Petal width, } x_4 \end{matrix} \tag{62}
$$

$$
\begin{matrix} \\ \\ \\ \\ \text{Sepal length, } x_1 \end{matrix} \quad \underbrace{}_{\text{Sepal width, } x_2} \quad \begin{matrix} \\ \\ \\ \\ \text{Petal length, } x_3 \end{matrix} \quad \begin{matrix} \\ \\ \\ \\ \text{Petal width, } x_4 \end{matrix}
$$

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 x_1 和花萼宽度 x_2 线性负相关，花瓣长度 x_3 和花萼宽度 x_2 线性负相关，花瓣宽度 x_4 和花萼宽度 x_2 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。

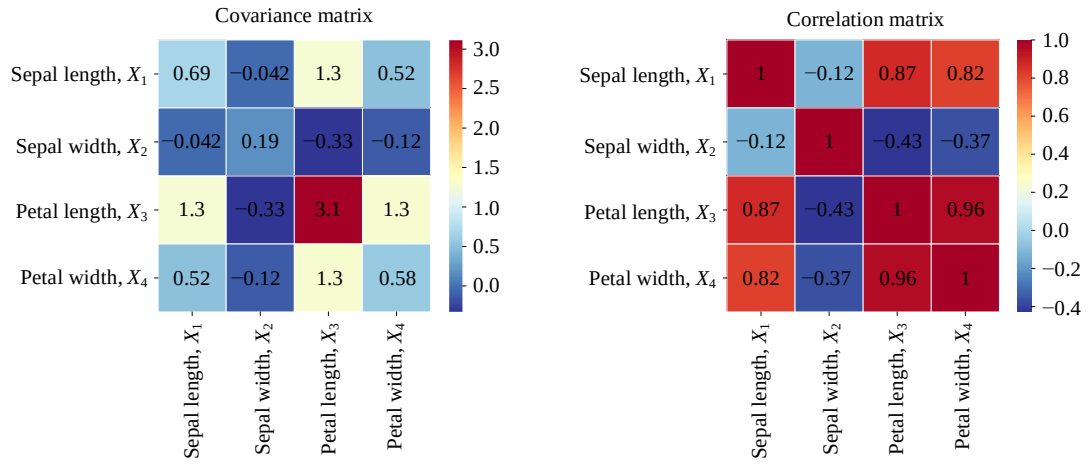


图 14. 协方差矩阵和相关性系数矩阵热图

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。

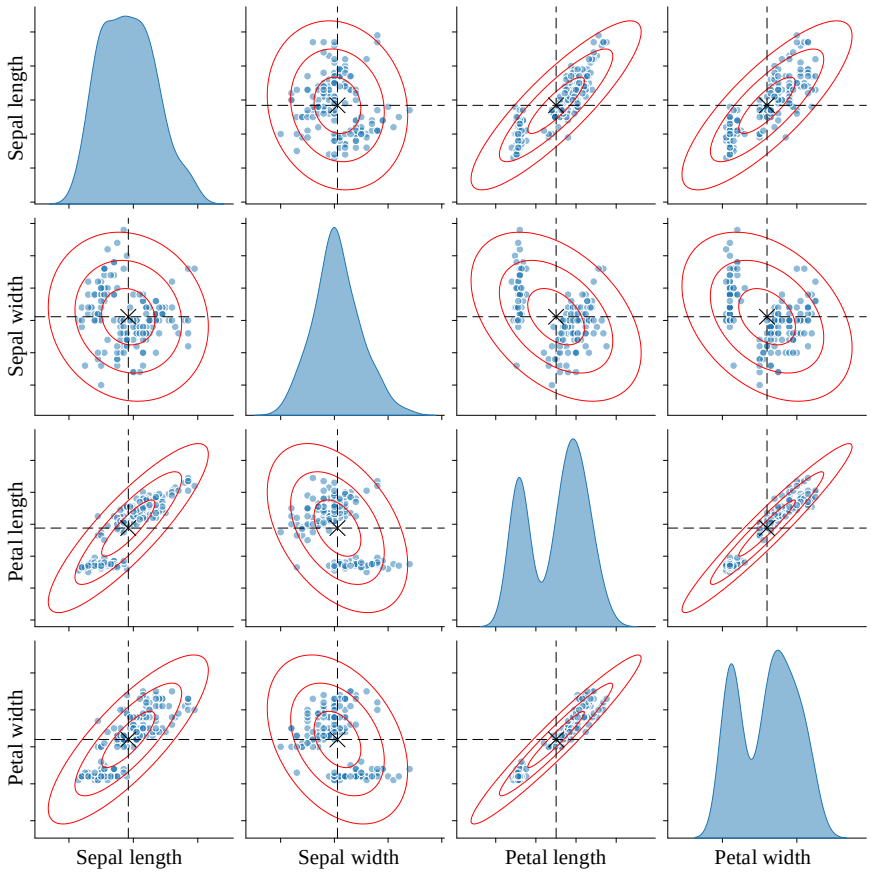


图 15. 协方差矩阵和椭圆的关系

如前文所述，鸢尾花数据分为三类。任意一簇 C_k 样本的分布也对应各自的协方差矩阵 $\boldsymbol{\Sigma}_k$（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。

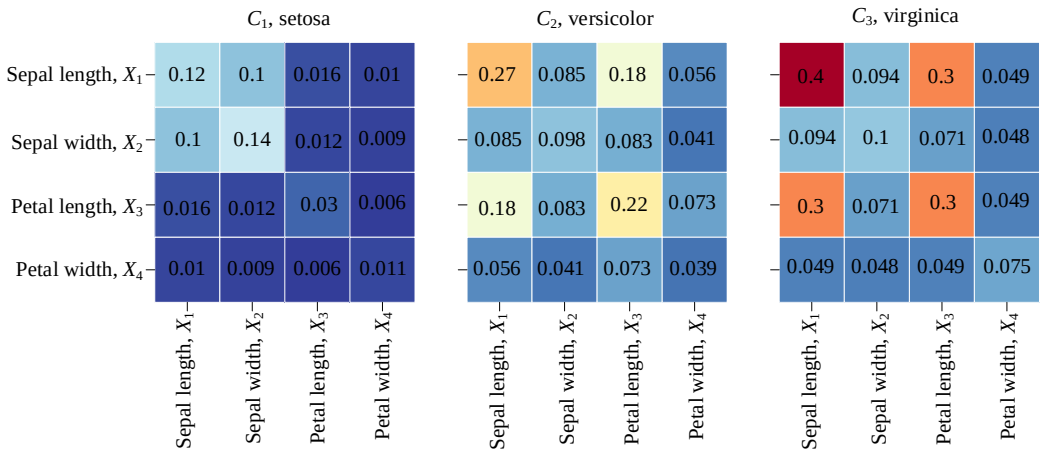


图 16. 协方差矩阵热图，考虑分类

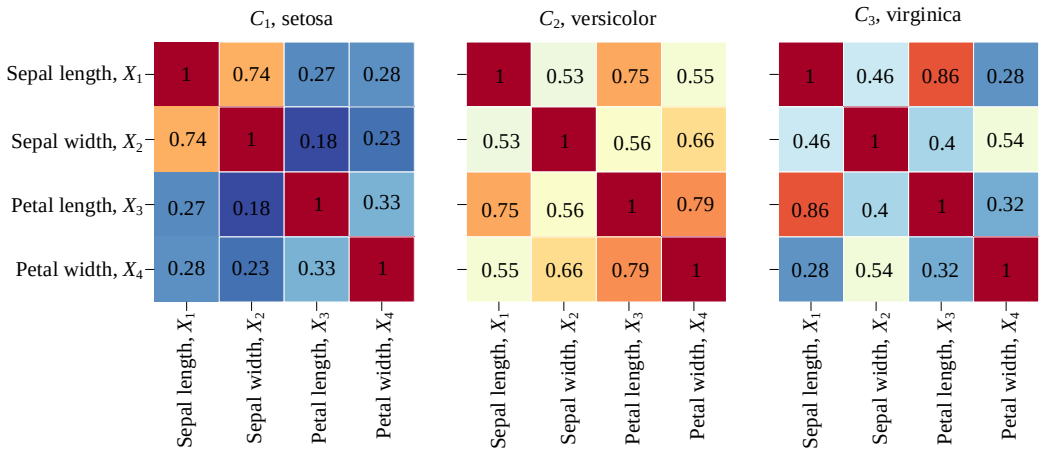


图 17. 相关性系数矩阵热图，考虑分类

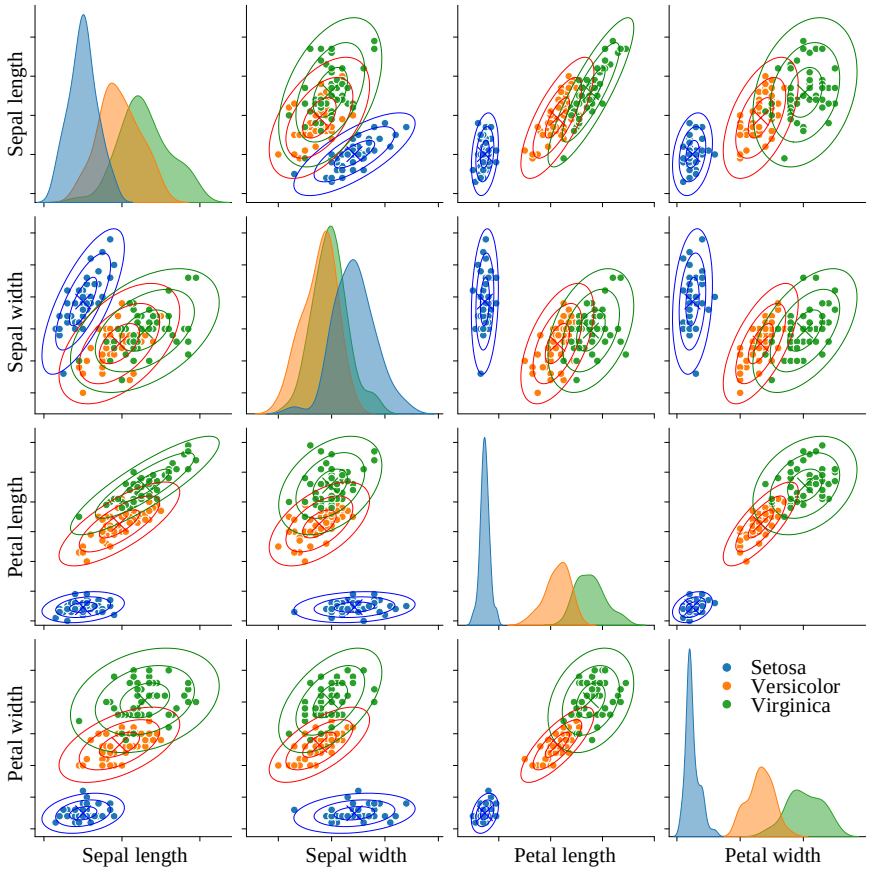
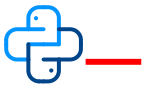


图 18. 协方差矩阵和椭圆的关系，考虑分类



Bk4_Ch22_01.py 中 Bk4_Ch22_01_D 部分绘制图 14、图 16、图 17 这几幅热图。

```
# Bk4_Ch22_01_D

#%% covariance matrix

SIGMA = X_df.cov()

fig, axs = plt.subplots()

h = sns.heatmap(SIGMA,cmap='RdBu_r', linewidths=.05, annot = True)
h.set_aspect("equal")
h.set_title('$\Sigma$')


#%% correlation matrix

RHO = X_df.corr()

fig, axs = plt.subplots()

h = sns.heatmap(RHO,cmap='RdBu_r', linewidths=.05, annot = True)
h.set_aspect("equal")
h.set_title('$\u03A1$')


#%% compare covariance matrices

f,(ax1,ax2,ax3) = plt.subplots(1,3,sharey=True)

g1 = sns.heatmap(X_df[y==0].cov(),cmap="RdYlBu_r",
                 annot=True,cbar=False,ax=ax1,square=True,
                 vmax = 0.4, vmin = 0)
ax1.set_title('Y = 0, setosa')

g2 = sns.heatmap(X_df[y==1].cov(),cmap="RdYlBu_r",
                 annot=True,cbar=False,ax=ax2,square=True,
                 vmax = 0.4, vmin = 0)
ax2.set_title('Y = 1, versicolor')

g3 = sns.heatmap(X_df[y==2].cov(),cmap="RdYlBu_r",
                 annot=True,cbar=False,ax=ax3,square=True,
                 vmax = 0.4, vmin = 0)
ax3.set_title('Y = 2, virginica')

#%% compare correlation matrices

f,(ax1,ax2,ax3) = plt.subplots(1,3,sharey=True)

g1 = sns.heatmap(X_df[y==0].corr(),cmap="RdYlBu_r",
                 annot=True,cbar=False,ax=ax1,square=True,
                 vmax = 1, vmin = 0.15)
ax1.set_title('Y = 0, setosa')

g2 = sns.heatmap(X_df[y==1].corr(),cmap="RdYlBu_r",
                 annot=True,cbar=False,ax=ax2,square=True,
                 vmax = 1, vmin = 0.15)
ax2.set_title('Y = 1, versicolor')

g3 = sns.heatmap(X_df[y==2].corr(),cmap="RdYlBu_r",
                 annot=True,cbar=False,ax=ax3,square=True,
                 vmax = 1, vmin = 0.15)
ax3.set_title('Y = 2, virginica')
```

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、向量、矩阵、统计之间的联系。

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 $\boldsymbol{Z}_X$。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。

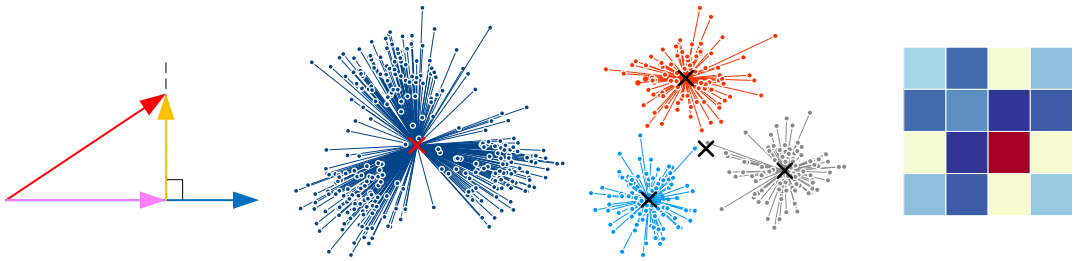


图 19. 总结本章重要内容的四副图

下一章正式进入本书的结束之旅——数据三部曲。