

22

Statistics Meet Linear Algebra

数据与统计

从线性代数运算视角看统计



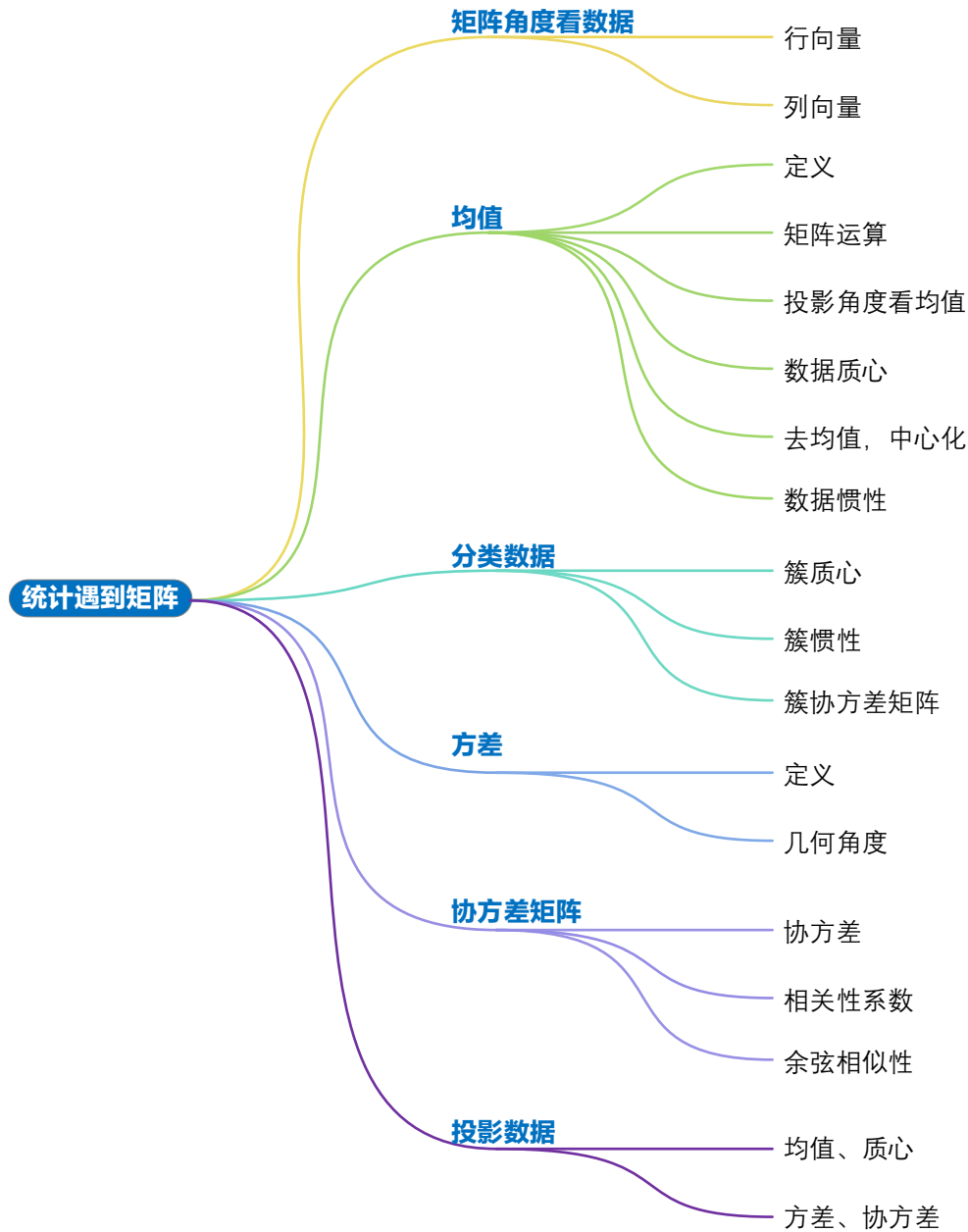
毫无争议的是，人类无法毫无错误地判断事物的真伪，我们能做就是遵循更大的可能性。

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 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 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。

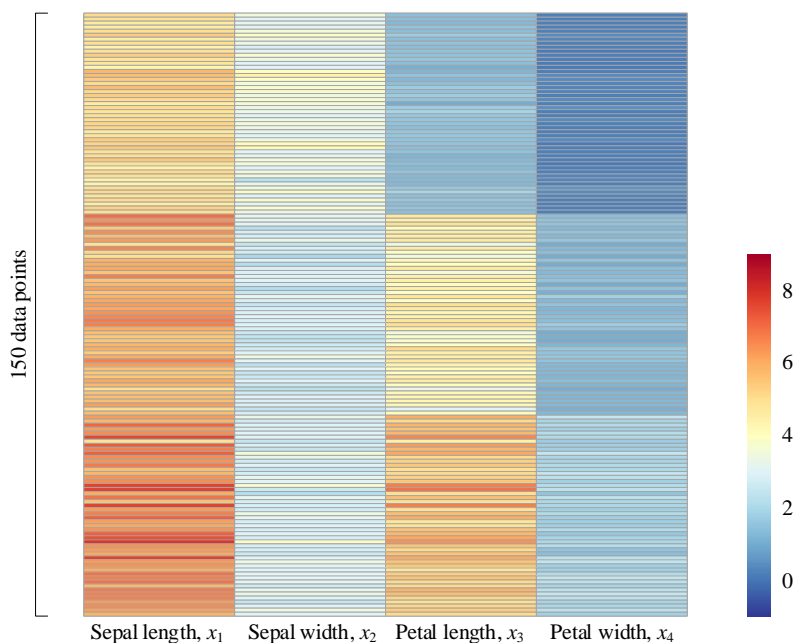


图 1. 鸢尾花数据，原始数据矩阵 X



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
```

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

```

# 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 \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：

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

回顾求解随机变量 X_j 的期望值 (均值) $E(X_j)$ 运算：

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

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

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

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

(3) 左乘 n 可以得到如下等式：

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

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

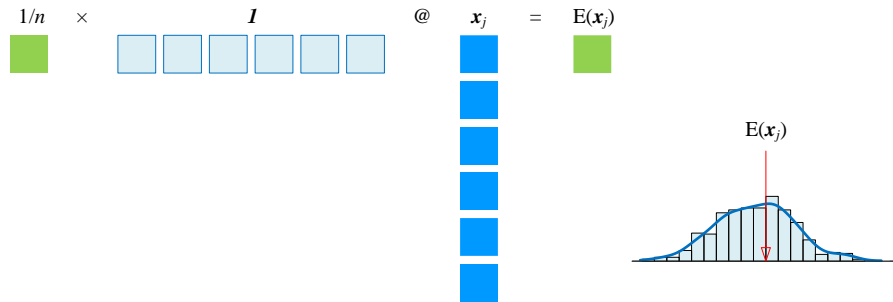


图 2. 计算 \mathbf{x}_j 期望值/均值

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：

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

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：

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

这部分垂直于 \mathbf{I} ，也就是：

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

也就是说，均值作为一个统计量，它能解释 \mathbf{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， $\mathbf{x}_j - \mathbf{E}(\mathbf{x}_j)\mathbf{I}$ ，将在标准差（方差平方根）中加以解释。

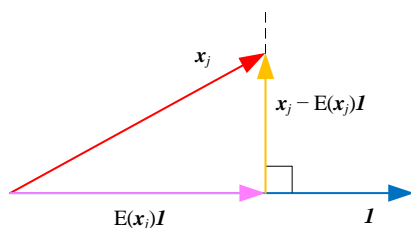


图 3. 投影角度看期望值

两个极端例子

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，

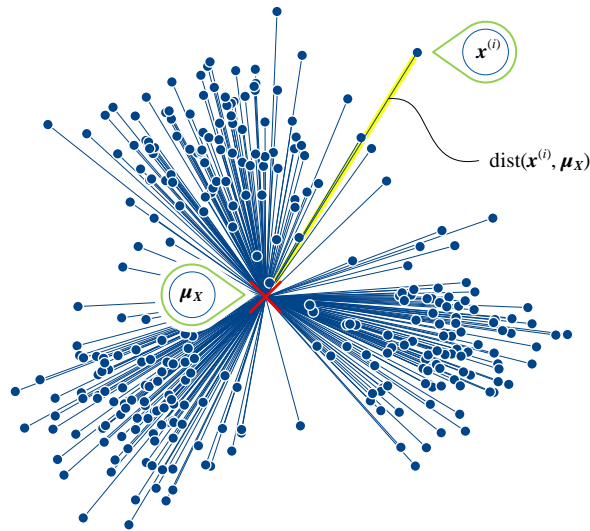
$$\mathbf{I}^T \mathbf{x}_j = 0 \quad (10)$$

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

对于最小二乘法线性回归， $\mathbf{x}_j - \mathbf{E}(\mathbf{x}_j)\mathbf{I}$ 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 \mathbf{X} 的质心位置。

图 4. 平面上数据矩阵 X 质心位置

列向量

X 样本数据的质心 μ_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， μ_X 被定义为列向量。

在多元高斯分布中，我们会用列向量 μ_X ：

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

上式中，几何角度来看， $x - \mu_X$ 相当于“平移”。

前文介绍过， μ_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。

举个例子，鸢尾花数据质心位置：

$$\mu_x = \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \quad (14)$$

整理 (13) 上式得到两个等式：

$$\begin{cases} X^T I = n\mu_x \\ I^T X = n(\mu_x)^T \end{cases} \quad (15)$$

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 $E(X)$ 为行向量，即：

$$\begin{aligned} E(X) &= [E(x_1) \ E(x_2) \ \cdots \ E(x_D)] \\ &= [\mu_1 \ \mu_2 \ \cdots \ \mu_D] \\ &= (\mu_x)^T = \frac{I^T X}{n} \end{aligned} \quad (16)$$

整理 (16)，可以得到：

$$I^T X = nE(X) \quad (17)$$

图 5 所示为计算质心示意图，以及 $E(X)$ 和 μ_x 之间关系。

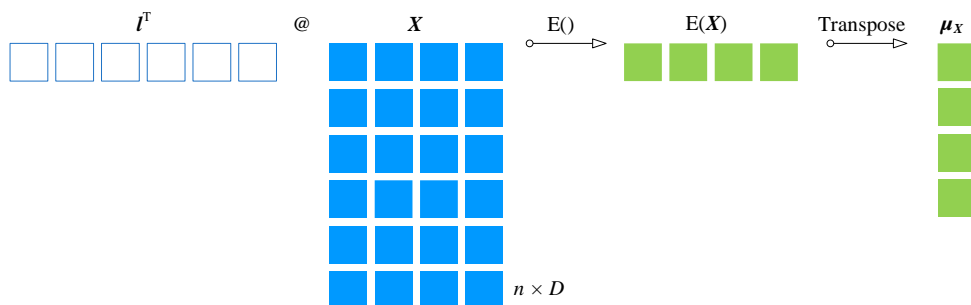


图 5. 计算 X 样本数据的质心 μ_x

22.4 中心化：平移

中心化、去均值

数据矩阵 X 中第 j 特征特征数据 x_j 减去均值 μ_j ，对应的矩阵运算为：

$$x_j - \mathbf{1}\mu_j = x_j - \frac{1}{n}\mathbf{1}\mathbf{1}^T x_j = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)x_j \quad (18)$$

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

而数据矩阵 X 中每一列数据 x_j 分别减去对应本列均值 μ_j ，对应矩阵运算为：

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

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 X ，即用原始数据矩阵 X 减去行向量 $E(X)$ ：

$$X_c = X - E(X) \quad (20)$$

图 6 所示为鸢尾花数据去均值后热图。中心化后，数据 X_c 质心位于原点。

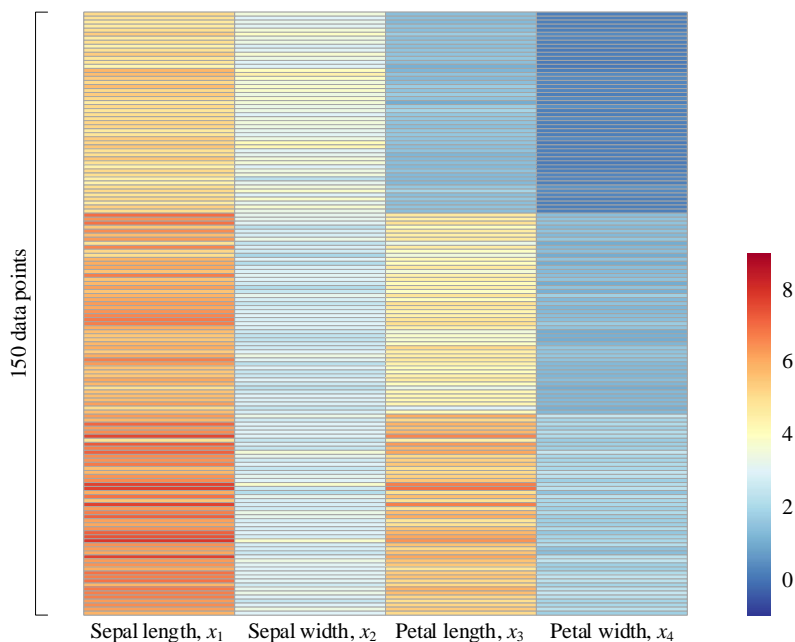


图 6. 鸢尾花数据矩阵 X ，去均值

标准化：平移 + 缩放

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

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：

$$\mathbf{Z}_X = \mathbf{X}_c \mathbf{S}^{-1} = (\mathbf{X} - \mathbf{E}(\mathbf{X})) \mathbf{S}^{-1} \quad (21)$$

其中，缩放矩阵 \mathbf{S} 为：

$$\mathbf{S} = \text{diag}(\text{diag}(\boldsymbol{\Sigma}))^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix} \quad (22)$$

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：

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

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。

(23) 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 `trace()` 可以方便得到 SSD 结果：

$$\text{SSD}(\mathbf{X}) = \text{trace}(\mathbf{X}_c^T \mathbf{X}_c) = \text{trace}\left((\mathbf{X} - \mathbf{E}(\mathbf{X}))^T (\mathbf{X} - \mathbf{E}(\mathbf{X}))\right) \quad (24)$$



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。

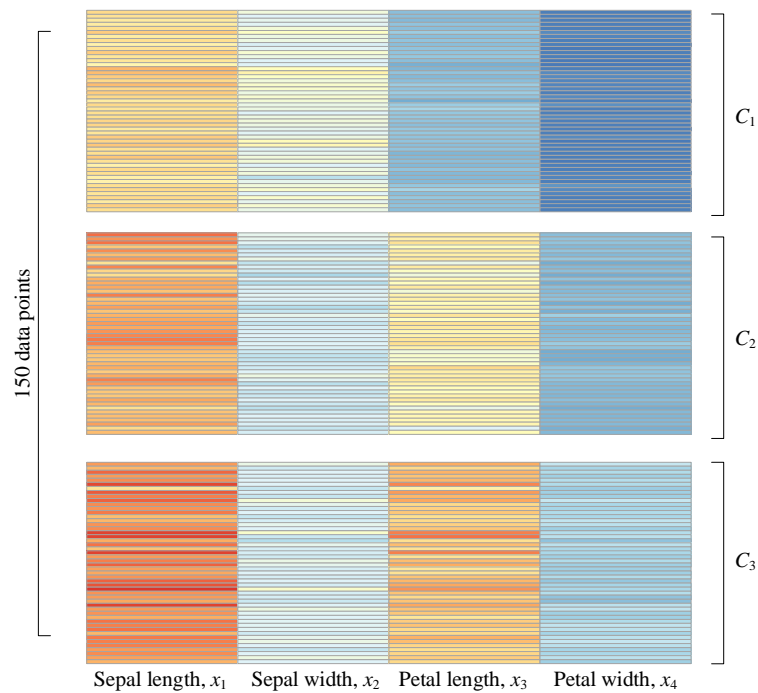


图 7. 鸢尾花数据分为三类

簇质心

类似 μ_X ，任意一类标签为 C_k 样本数据的簇质心 μ_k ，定义如下：

$$\mu_k = \frac{1}{\text{count}(C_k)} \sum_{i \in C_k} \mathbf{x}^{(i)} \quad (25)$$

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 $\mathbf{x}^{(i)} (i \in C_k)$ ，求解其各个特征位置平均值，构造一个新的列向量 μ_k 。图 8 所示为样本数据质心 μ_X ，和三类数据各自的质心 μ_1 、 μ_2 和 μ_3 之间的关系。

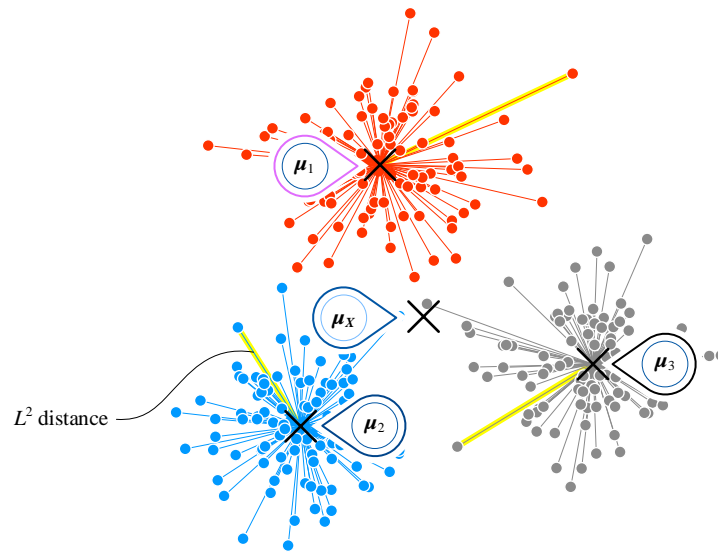


图 8. 样本数据质心 μ_X ，和三类数据各自的质心 μ_1 、 μ_2 和 μ_3

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

数据点有两特征，具体坐标值如下：

$$\mathbf{x}^{(2)} = [2 \ 3]^T, \mathbf{x}^{(5)} = [3 \ 1]^T, \mathbf{x}^{(6)} = [-2 \ 2]^T, \mathbf{x}^{(9)} = [1 \ 6]^T \quad (26)$$

则标签为 C_1 簇质心位置为 $[1, 3]^T$ ，具体运算过程如下：

$$\begin{aligned} \mu_{C_1} &= \frac{1}{\text{count}(C_1)} \sum_{i \in C_1} \mathbf{x}^{(i)} = \frac{1}{\text{count}(C_1)} (\mathbf{x}^{(2)} + \mathbf{x}^{(5)} + \mathbf{x}^{(6)} + \mathbf{x}^{(9)}) \\ &= \frac{1}{4} ([2 \ 3]^T + [3 \ 1]^T + [-2 \ 2]^T + [1 \ 6]^T) \\ &= [1 \ 3]^T \end{aligned} \quad (27)$$

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 \times 代表鸢尾花的簇质心位置。

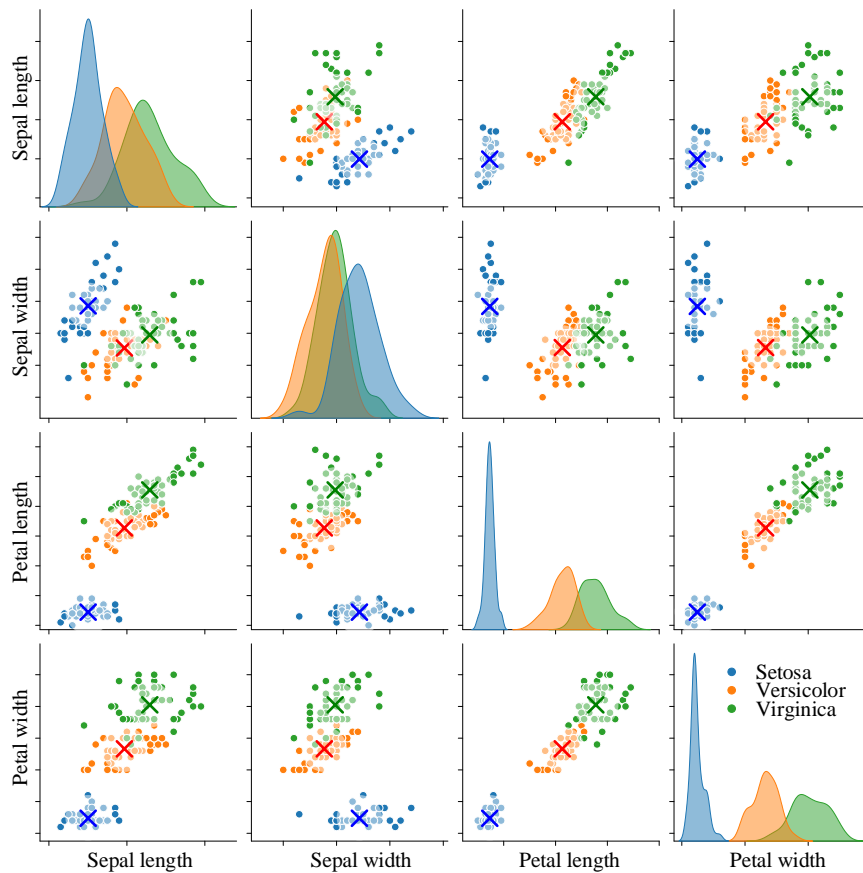


图 9. 鸢尾花数据簇质心位置

簇惯量

簇惯量 (cluster inertia) 定义如下：

$$\text{SSD}(C_k) = \sum_{i \in C_k} \|\mathbf{x}^{(i)} - \boldsymbol{\mu}_k\|_2^2 \quad (28)$$

(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：

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

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

对于样本来说，随机变量 X 方差的计算式为：

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

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

$$\begin{aligned} \text{var}(X) &= E\left(\left(X - E(X)\right)^2\right) \\ &= E\left(X^2 - 2XE(X) + E(X)^2\right) \\ &= E(X^2) - E(X)^2 \end{aligned} \quad (31)$$

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

$$\text{var}(\mathbf{x}) = E(\mathbf{x} \odot \mathbf{x}) - E(\mathbf{x})^2 \quad (32)$$

其中， \odot 为向量逐项积运算符。

进一步展开 (32) 得到：

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

根据 (32)，还可以得到如下等式：

$$E(\mathbf{x} \odot \mathbf{x}) = \text{var}(\mathbf{x}) + E(\mathbf{x})^2 \quad (34)$$

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

$$\mathbf{x}^T \mathbf{x} = n\left(\text{var}(\mathbf{x}) + E(\mathbf{x})^2\right) \quad (35)$$

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， \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 n 倍方差。

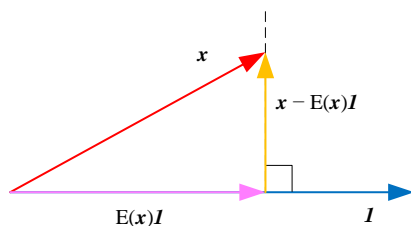


图 10. 投影角度看方差和标准差

对于数据矩阵 X 而言，第 j 列数据 x_j 的方差有几种不同表达方式：

$$\text{var}(X_j) = \text{var}(x_j) = \sigma_j^2 = \sigma_{j,j} \quad (36)$$

计算鸢尾花数据 X 每一列标准差，以行向量表达：

$$\sigma_X = \begin{bmatrix} 0.825 & 0.434 & 1.759 & 0.759 \\ \text{Sepal length, } x_1 & \text{Sepal width, } x_2 & \text{Petal length, } x_3 & \text{Petal width, } x_4 \end{bmatrix} \quad (37)$$

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。

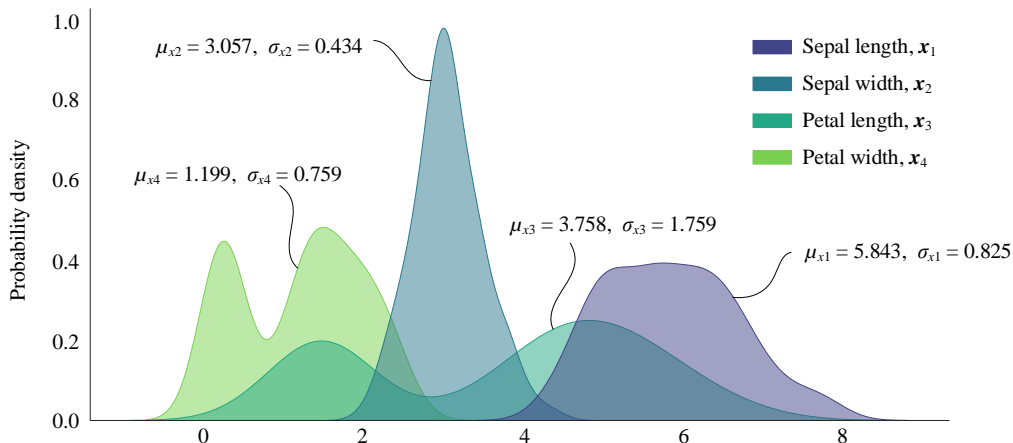


图 11. 鸢尾花数据四个特征上分布



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

$$\begin{aligned}\text{cov}(\mathbf{x}, \mathbf{y}) &= \mathbb{E}[(\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))] \\ &= \mathbb{E}[\mathbf{x} \odot \mathbf{y} - \mathbf{x} \mathbb{E}(\mathbf{y}) - \mathbb{E}(\mathbf{x}) \mathbf{y} + \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y})] \\ &= \mathbb{E}(\mathbf{x} \odot \mathbf{y}) - \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y}) - \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y}) + \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y}) \\ &= \mathbb{E}(\mathbf{x} \odot \mathbf{y}) - \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y})\end{aligned}\quad (38)$$

上式中 $\mathbb{E}(\mathbf{x} \odot \mathbf{y})$ 是 \mathbf{x} 向量和 \mathbf{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：

$$\mathbb{E}(\mathbf{x} \odot \mathbf{y}) = \text{cov}(\mathbf{x}, \mathbf{y}) + \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y}) \quad (39)$$

用向量内积方式来写，协方差为：

$$\text{cov}(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - \mathbb{E}(\mathbf{x})) \cdot (\mathbf{y} - \mathbb{E}(\mathbf{y}))}{n} \quad (40)$$

将具体数值代入 (38)，并整理得到：

$$\begin{aligned}\text{cov}(\mathbf{x}, \mathbf{y}) &= \mathbb{E}(\mathbf{x} \odot \mathbf{y}) - \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y}) \\ &= \frac{\mathbf{x}^T \mathbf{y}}{n} - \frac{1}{n} \mathbf{x}^T \mathbf{I} \frac{1}{n} \mathbf{y}^T \mathbf{I} = \frac{n \mathbf{x}^T \mathbf{y} - \mathbf{x}^T \mathbf{I} \mathbf{y}^T \mathbf{I}}{n^2} \\ &= \frac{n \left(\sum_{i=1}^n x_i y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n^2}\end{aligned}\quad (41)$$

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

$$\mathbf{y}^T \mathbf{x} = \mathbf{x}^T \mathbf{y} = n(\text{cov}(\mathbf{x}, \mathbf{y}) + \mathbb{E}(\mathbf{x}) \mathbb{E}(\mathbf{y})) \quad (42)$$

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

$$\text{cov}(\mathbf{X}_i, \mathbf{X}_j) = \text{cov}(\mathbf{x}_i, \mathbf{x}_j) = \rho_{i,j} \sigma_i \sigma_j = \sigma_{i,j} \quad (43)$$

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

$$\begin{aligned}\text{cov}(\mathbf{x}, \mathbf{I}) &= \mathbf{E}(\mathbf{x} \odot \mathbf{I}) - \mathbf{E}(\mathbf{x})\mathbf{E}(\mathbf{I}) \\ &= \mathbf{E}(\mathbf{x}) - \mathbf{E}(\mathbf{x}) \\ &= \mathbf{0}\end{aligned}\quad (44)$$

相关性系数

随机变量 X 和 Y 相关性系数的定义为：

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

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

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x} - \mathbf{E}(\mathbf{x})) \cdot (\mathbf{y} - \mathbf{E}(\mathbf{y}))}{\|\mathbf{x} - \mathbf{E}(\mathbf{x})\| \|\mathbf{y} - \mathbf{E}(\mathbf{y})\|} = \left(\frac{\mathbf{x} - \mathbf{E}(\mathbf{x})}{\|\mathbf{x} - \mathbf{E}(\mathbf{x})\|} \right) \cdot \left(\frac{\mathbf{y} - \mathbf{E}(\mathbf{y})}{\|\mathbf{y} - \mathbf{E}(\mathbf{y})\|} \right) \quad (46)$$

相信大家已经在上式中，看到“平移”和“缩放”。也就是说，我们将 $[\mathbf{x}, \mathbf{y}]$ 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：

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

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：

$$\begin{aligned}\mathbf{x} \cdot \mathbf{y} &= \mathbf{y} \cdot \mathbf{x} \\ \text{cov}(X, Y) &= \text{cov}(Y, X)\end{aligned}\quad (48)$$

向量的模类似标准差：

$$\begin{aligned}\|\mathbf{x}\| &= \sqrt{\mathbf{x} \cdot \mathbf{x}} \\ \sigma_X &= \sqrt{\text{var}(X)} = \sqrt{\text{cov}(X, X)}\end{aligned}\quad (49)$$

向量之间夹角余弦值类似线性相关性系数：

$$\begin{aligned}\cos \theta &= \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} \\ \rho_{X,Y} = \text{corr}(X, Y) &= \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbf{E}((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y}\end{aligned}\quad (50)$$

(50) 可以整理成如下等式：

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

此外，余弦定理可以用在向量内积和协方差上：

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

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。

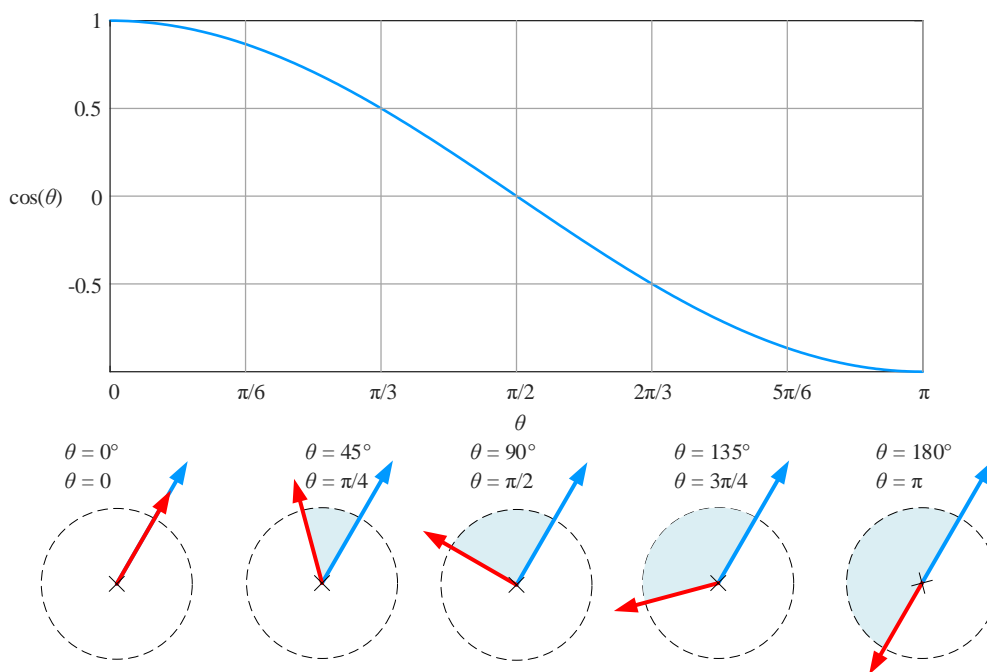


图 12. 余弦相似度

22.8 协方差矩阵和相关性系数矩阵

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

协方差矩阵

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

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

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 \mathbf{X} 计算得到：

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

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 $\mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x}$ 。

对 $\boldsymbol{\Sigma}$ 进行特征值分解，得到：

$$\boldsymbol{\Sigma} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \quad (55)$$

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

$$\begin{aligned} \mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x} &= \mathbf{x}^T \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T \mathbf{x} \\ &= \begin{pmatrix} \mathbf{V}^T \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \boldsymbol{\Lambda} \begin{pmatrix} \mathbf{V}^T \mathbf{x} \\ \mathbf{y} \end{pmatrix} \\ &= \mathbf{y}^T \boldsymbol{\Lambda} \mathbf{y} \\ &= \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_D y_D^2 = \sum_{j=1}^D \lambda_j y_j^2 \end{aligned} \quad (56)$$

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

$$\mathbf{x}^T \boldsymbol{\Sigma} \mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \sigma_{1,1} x_1^2 + (\sigma_{1,2} + \sigma_{2,1}) x_1 x_2 + \sigma_{2,2} x_2^2 \quad (57)$$

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

$$\mathbf{y}^T \mathbf{A} \mathbf{y} = \begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \lambda_1 y_1^2 + \lambda_2 y_2^2 \quad (58)$$

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

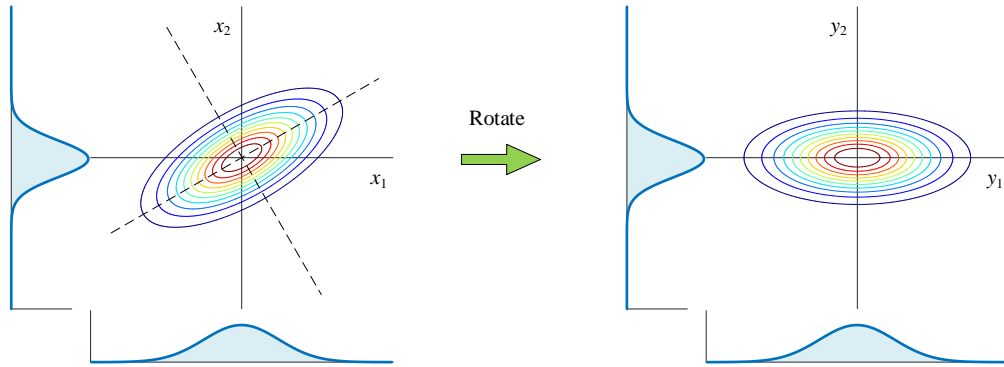


图 13. 旋转椭圆到正椭圆

相关性系数矩阵

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

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix} \quad (59)$$

\mathbf{P} 和 $\mathbf{\Sigma}$ 的关系为：

$$\mathbf{\Sigma} = \mathbf{S} \mathbf{P} \mathbf{S}^T \quad (60)$$

\mathbf{S} 就是 (22) 定义的缩放矩阵， \mathbf{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 $\mathbf{\Sigma}$ 为：

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

鸢尾花数据的相关性系数矩阵 \mathbf{P} 为：

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

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

图 14 所示为 Σ 和 \mathbf{P} 的热图。观察相关性系数矩阵 \mathbf{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。

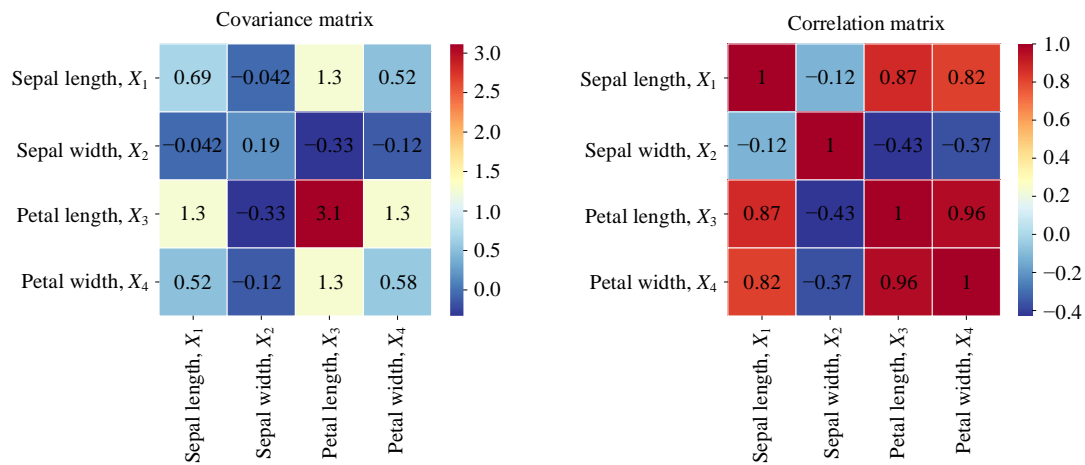


图 14. 协方差矩阵和相关性系数矩阵热图

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。

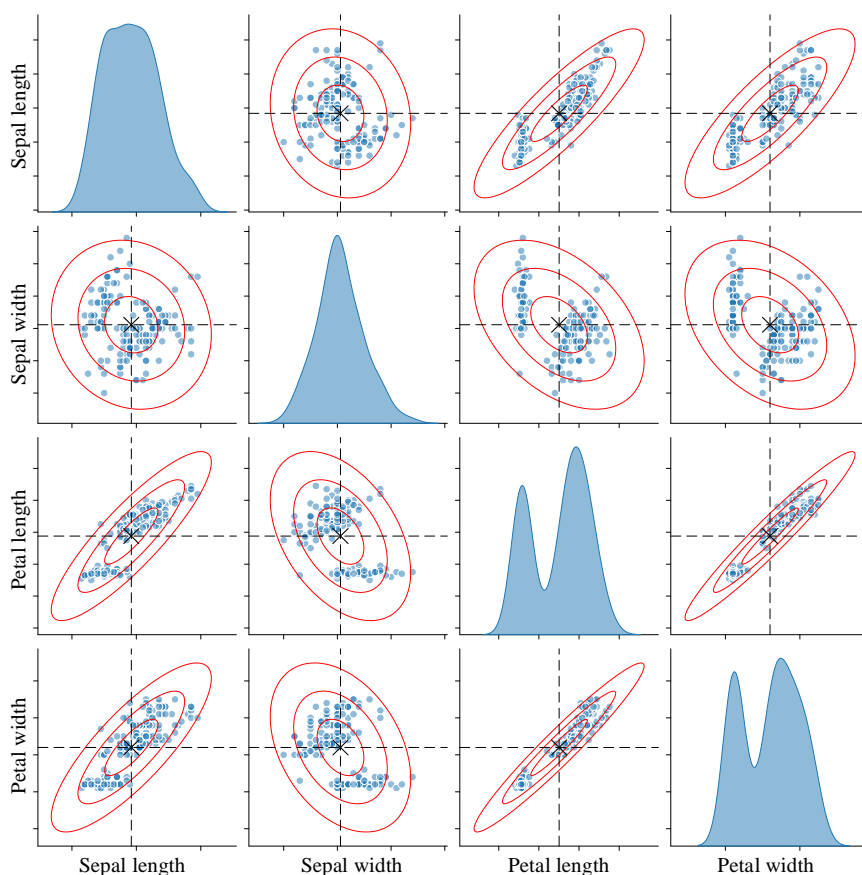


图 15. 协方差矩阵和椭圆的关系

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 Σ_k (如图 16) 和相关系数矩阵 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。

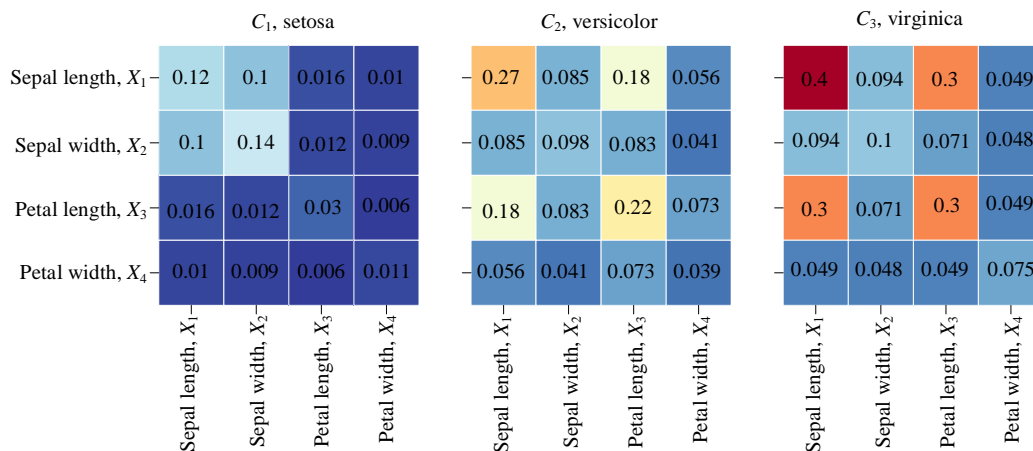


图 16. 协方差矩阵热图，考虑分类

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

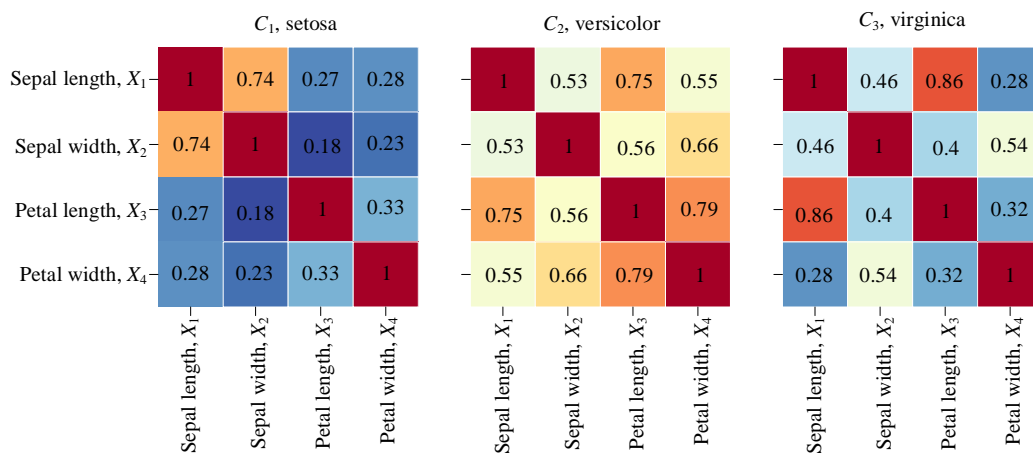


图 17. 相关性系数矩阵热图，考虑分类

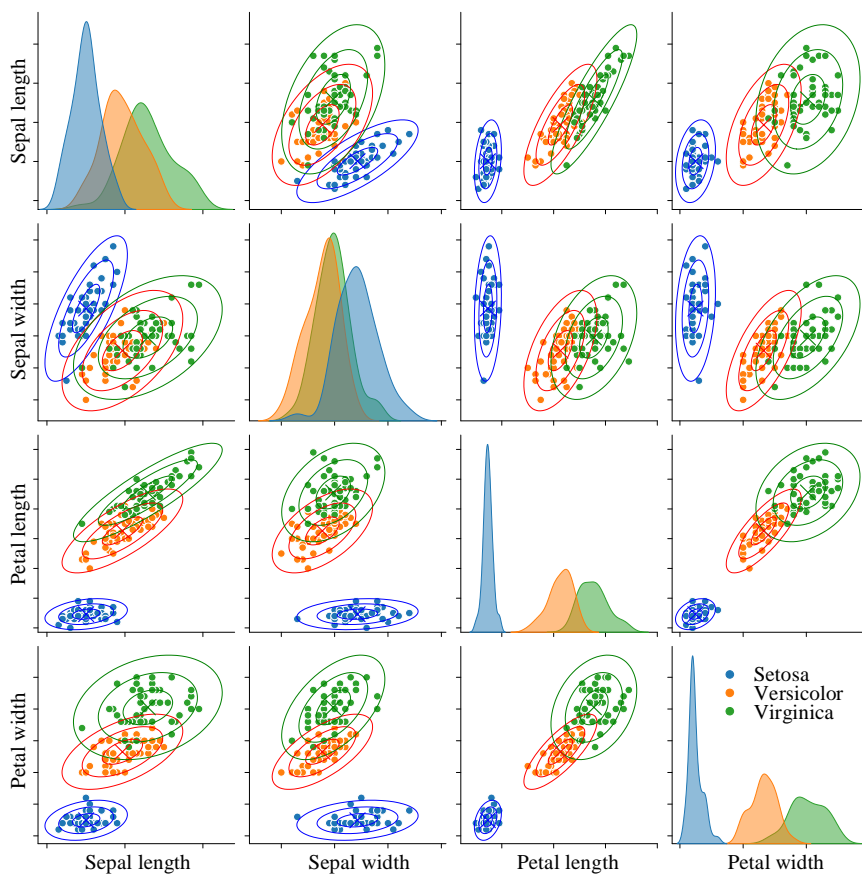


图 18. 协方差矩阵和椭圆的关系，考虑分类

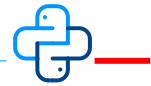
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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



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('$\rho$')

### 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')

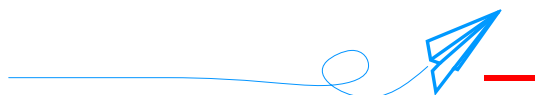
### compare correlation matrices

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 X 形状相同的数据矩阵——中心化数据矩阵 X_c 、标准化数据矩阵 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。