作者	**生姜 DrGinger**
脚本	**生姜 DrGinger**
视频	**崔崔 CuiCui**
开源学习资源	**https://github.com/Visualize-ML**
平台	**https://www.youtube.com/@DrGinger_Jiang** **https://space.bilibili.com/3546865719052873** **https://space.bilibili.com/513194466**

12 Dimensionality Reduction 降维

将多维数据投影到低维空间，揭示数据内在结构

在本章中，我们将从最基础的数据矩阵出发，逐步展开对主成分分析的深入理解。本章有一个特殊的格拉姆矩阵——协方差矩阵——扮演核心角色。我们将利用上一章介绍的特征值分解探索协方差矩阵，从而了解数据结构，并实现对数据的降维、压缩与特征提取。

12.1 数据矩阵

本节你将掌握的核心技能：

- 数据矩阵的行表示样本、列表示特征。
- 通过矩阵乘法求取每列特征均值及整个数据质心。
- 去均值操作，使数据的质心移动至原点，方便后续协方差分析。
- 协方差矩阵是中心化数据格拉姆矩阵的一种，反映内积结构。
- 协方差矩阵，对角线上的方差和非对角线上的协方差。
- 标准化：各列减去均值再除以标准差，使数据无单位、均值为 0、标准差为 1。
- 线性相关系数是对协方差的归一化，不受量纲影响，便于特征比较。
- 线性相关性系数矩阵：标准化数据的协方差矩阵，也是格拉姆矩阵。

本节从数据矩阵开始一步步讲解围绕在数据矩阵周围的各种常见操作，比如质心计算、中心化、协方差矩阵计算、标准化、线性相关性系数矩阵计算等等。各种线性代数工具将大显身手！

值得注意的是，协方差矩阵、线性相关性系数矩阵可以看作是数据矩阵的两种特殊格拉姆矩阵。

数据矩阵

数据矩阵 (data matrix) 将现实世界中的各种数据整理成“行 × 列”形式的一种方式。

最直观的例子是电子表格，每一行代表一个**样本** (sample)，每一列代表一个特征。

例如，在一个关于学生成绩的表格中，每一行是一个学生，每一列可能是数学、语文、英语等科目的成绩。这就是一个典型的数据矩阵。

“鸢尾花书”一般会用矩阵 $\boldsymbol{X}$ (大写、斜体、粗体) 代表数据矩阵。

如图 1 (a) 所示，数据矩阵 $\boldsymbol{X}$ 有 n 行、D 列。

如图 1 (b) 所示，数据矩阵 $\boldsymbol{X}$ 的每一列代表一个特征 (或随机变量) 的样本数据；$\boldsymbol{X}$ 有 D 个特征，对应 $\boldsymbol{D}$ 列列向量。简单来说，随机变量是一个用来描述随机现象结果的变量。

如图 1 (c) 所示，数据矩阵 $\boldsymbol{X}$ 的每一行代表一个样本；$\boldsymbol{X}$ 有 n 行，即 n 个样本。

请预习 (复习) 如何计算样本数据均值、方差、标准差、协方差、线性相关性系数。

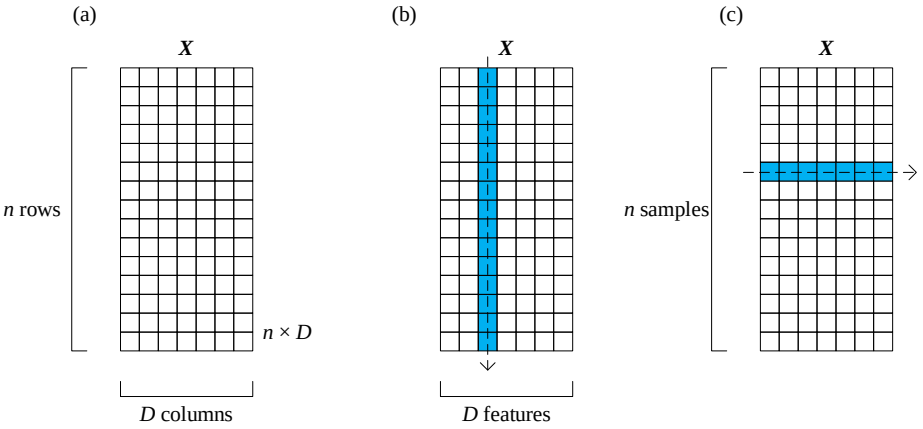

图 1. 数据矩阵

注意，本书代数变量会用小写、斜体 x、x_1、x_2、y、y_1、y_2 等；数据矩阵用大写、斜体、粗体 $\boldsymbol{X}$。数据列向量会用小写、斜体、粗体、下标数字 $\boldsymbol{x}$、$\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{y}$、$\boldsymbol{y}_1$、$\boldsymbol{y}_2$ 等；数据行向量 (样本) 用小写、斜体、粗体、上标数字 $\boldsymbol{x}^{(1)}$、$\boldsymbol{x}^{(2)}$、$\boldsymbol{y}^{(1)}$、$\boldsymbol{y}^{(2)}$；随机变量会用大写、斜体、下标数字 X、X_1、X_2、Y、Y_1、Y_2 等。

行向量、列向量

图 2 所示为鸢尾花数据前 4 列 (花萼长度、花萼宽度、花瓣长度、花瓣宽度) 特征的数据。

数据的每一行 (行向量) 代表一朵鸢尾花样本；比如，第一行行向量

$$\boldsymbol{x}^{(1)} = \begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \end{bmatrix} \tag{1}$$

⚠ 注意，图 2 中所有的数据单位为厘米 (cm)；图 2 没有考虑鸢尾花标签列。

图 2 中每一列向量代表一个特征，如花萼长度、花萼宽度、花瓣长度、花瓣宽度。

比如，第一列为花萼长度数据

$$\boldsymbol{x}_1 = \begin{bmatrix} 5.1 \\ 4.9 \\ 4.7 \\ \vdots \end{bmatrix} \tag{2}$$

这一列有 150 个数值，对应 150 朵鸢尾花样本。

基于这些数据可以进行各种统计运算，如计算均值、方差、标准差等等。而基于两组列向量数据可以计算协方差、线性相关性系数。这些都是本节后续要介绍的知识点。

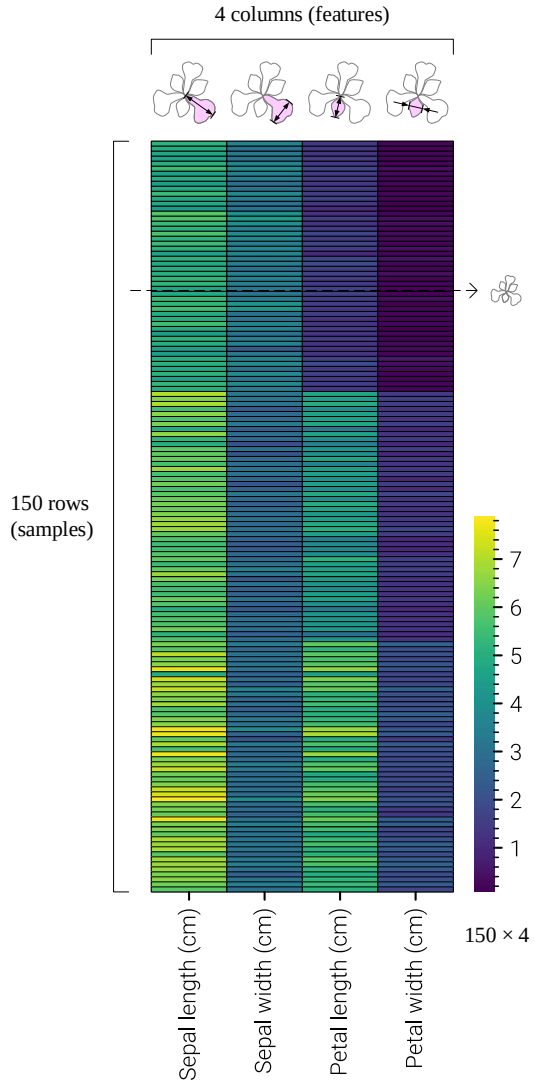

图 2. 鸢尾花数据矩阵热图，单位为厘米，没有展示最后标签列

本书第 10 章、第 1 节用平面散点图展示两个特征数据，第 10 章、第 3 节用三维空间散点图展示三个特征数据。对于图 2 所示 4 个特征数据，为了展示数据的分布情况，我们需要用成对散点图。

图 3 所示为鸢尾花数据绘制的成对散点图。

图 3 也相当于一个 4 × 4 矩阵，矩阵的每个元素是子图。

图 3 中非主对角线子图为平面散点图，大家对此应该很熟悉了。

图 3 中主对角线子图为**核密度估计图** (Kernel Density Estimation plot, KDE plot)；这种图是一种用平滑曲线近似数据分布的可视化工具，用于展示数据的概率密度分布，比直方图更平滑，有助于理解数据集中趋势和分布形态。

可以想象图 2 中数据相当于四维空间中散点，将它们分别投影到不同平面便得到图 3 子图。

比如，数据矩阵 $\boldsymbol{X}$ 向 x_1 轴投影相当于取出 $\boldsymbol{X}$ 的第一列 $\boldsymbol{x}_1$，对应矩阵乘法

$$\boldsymbol{X} @ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \cdots & \boldsymbol{x}_D \end{bmatrix} @ \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \boldsymbol{x}_1 \tag{3}$$

这些数据用来绘制图 2 第一行、第一列图像。

数据矩阵 $\boldsymbol{X}$ 向 x_1x_2 平面投影相当于取出 $\boldsymbol{X}$ 的前两列

$$\boldsymbol{X} @ \begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x}_3 & \cdots & \boldsymbol{x}_D \end{bmatrix} @ \begin{bmatrix} 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix} \tag{4}$$

这些数据用来绘制图 2 第二行、第一列图像；这个图像转置之后得到图 2 第一行、第二列图像。

相信大家已经发现图 3 相当于一个对称矩阵，我们实际上只需要主对角线上子图，以及主对角线以上、或以下六幅散点图。

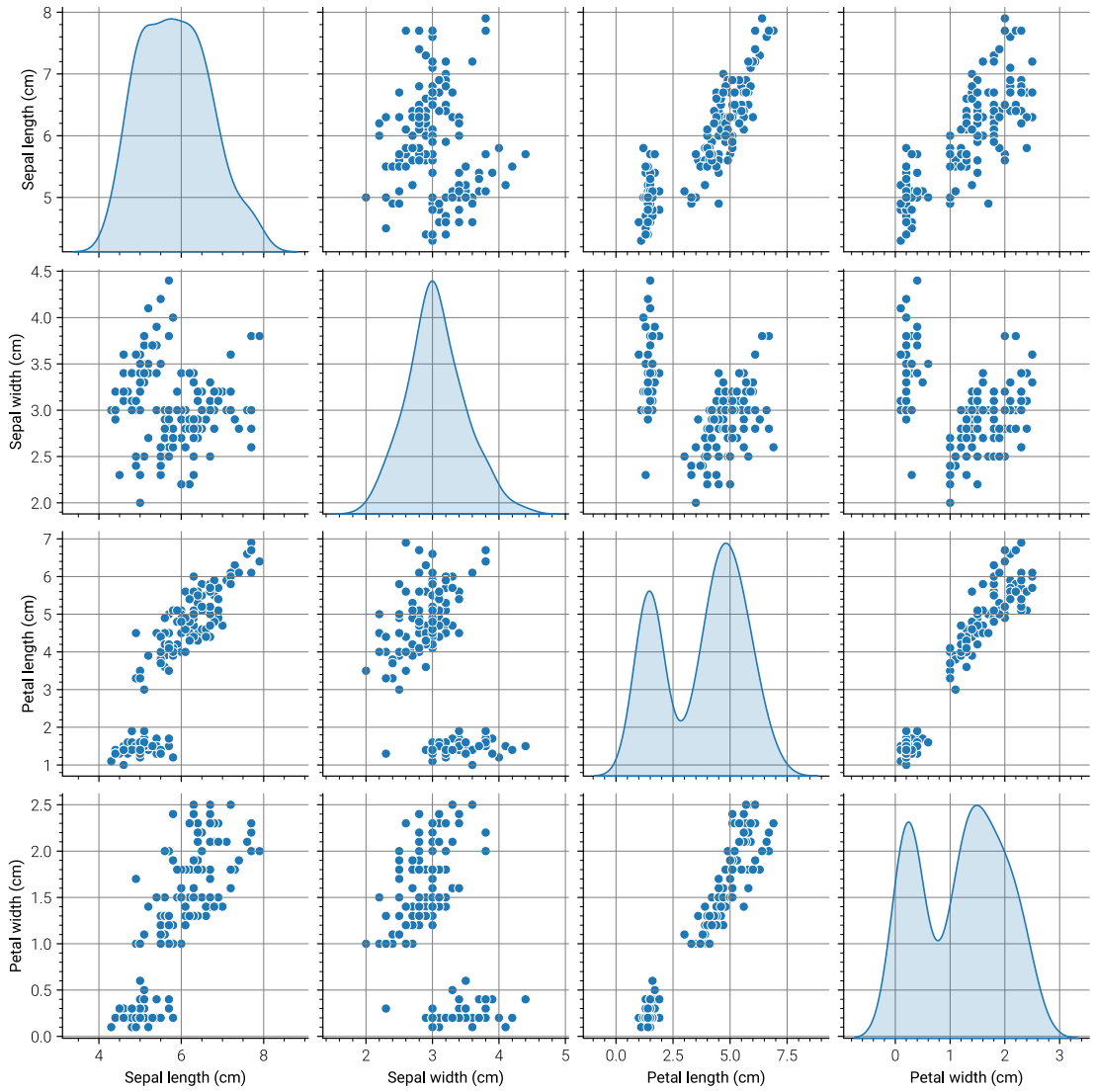

图 3. 鸢尾花数据成对散点图

均值、质心

在数据分析中，**均值** (mean, average) 是最常见的集中趋势度量，它表示一组数据的“平均”位置。

对于鸢尾花数据集，$\boldsymbol{x}_1$ 的均值 μ_1 就是花萼长度所有数据点值的总和除以数据个数 ($n = 150$)，对应矩阵乘法

$$\mu_1 = \frac{\boldsymbol{1}^{\mathrm{T}}\boldsymbol{x}_1}{n} = \frac{\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{1}}{n} = \frac{\boldsymbol{1}\cdot\boldsymbol{x}_1}{n} = \frac{\boldsymbol{x}_1\cdot\boldsymbol{1}}{n} = \frac{\sum_{i=1}^{n} x_{i,1}}{n} \tag{5}$$

注意，上式中全 1 列向量 $\boldsymbol{1}$ 有 n 行。

请大家用本节配套代码分别计算各个特征均值。

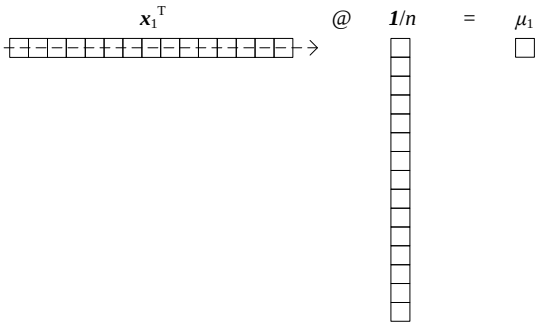

图 4. 计算 $\boldsymbol{x}_1$ 的均值

很容易计算得到鸢尾花四个特征的样本均值

$$\begin{cases} \mu_1 = 5.843 \\ \quad \text{Sepal length} \\ \mu_2 = 3.057 \\ \quad \text{Sepal width} \\ \mu_3 = 3.758 \\ \quad \text{Petal length} \\ \mu_4 = 1.199 \\ \quad \text{Petal width} \end{cases} \tag{6}$$

⚠ 注意，(6) 本书中质心 $\boldsymbol{\mu}$ 为列向量；均值的单位均为厘米，和样本数据单位一致。

如果 (5) 不除 n 的话，得到的是样本值求和，即

$$\sum_{i=1}^{n} x_{i,1} = \boldsymbol{1}^{\mathrm{T}} \boldsymbol{x}_1 = \boldsymbol{x}_1^{\mathrm{T}} \boldsymbol{1} = \boldsymbol{1} \cdot \boldsymbol{x}_1 = \boldsymbol{x}_1 \cdot \boldsymbol{1} = n\mu_1 \tag{7}$$

质心 (centroid)，$\boldsymbol{\mu}$，是多元数据中的一种位置描述，它是所有样本数据点的几何中心，即

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} \tag{8}$$

在几何上，质心是能够使所有样本点“平衡”的那个点。

⚠ 注意，本书中质心 $\boldsymbol{\mu}$ 为列向量。

如图 5 所示，数据矩阵 $\boldsymbol{X}$ 的质心可以通过如下矩阵乘法得到

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} = \frac{\boldsymbol{X}^{\mathrm{T}} \boldsymbol{1}}{n} \tag{9}$$

⚠ 注意，上式中全 1 列向量 $\boldsymbol{1}$ 有 n 行。

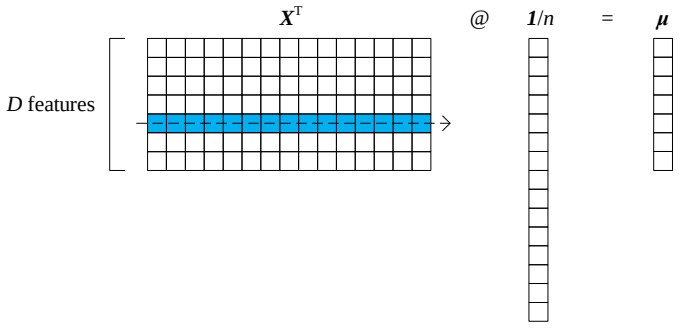

图 5. 计算数据矩阵 $\boldsymbol{X}$ 的质心

鸢尾花数据的质心为

$$\boldsymbol{\mu} = \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} \tag{5}$$

图 6 中，我们把质心位置用红色 × 画在成对散点图上。

质心 $\boldsymbol{\mu}$ 向 x_1 轴投影，便得到 μ_1

$$\mu_1 = \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{e}_1 = \boldsymbol{e}_1^{\mathrm{T}}\boldsymbol{\mu} = \boldsymbol{\mu} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} @ \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} = 5.843 \tag{10}$$

请大家思考如何用质心计算其他几个特征的均值。

质心 $\boldsymbol{\mu}$ 向 x_1x_2 平面投影，便得到 $[\mu_1, \mu_2]^{\mathrm{T}}$，对应如下矩阵乘法

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} @ \begin{bmatrix} 5.843 \\ 3.057 \\ 3.758 \\ 1.199 \end{bmatrix} = \begin{bmatrix} 5.843 \\ 3.057 \end{bmatrix} \tag{11}$$

这个质心投影对应图 6 第二行、第一列子图中 × 位置。

请大家思考如何计算图 6 其他几个子图质心投影位置。

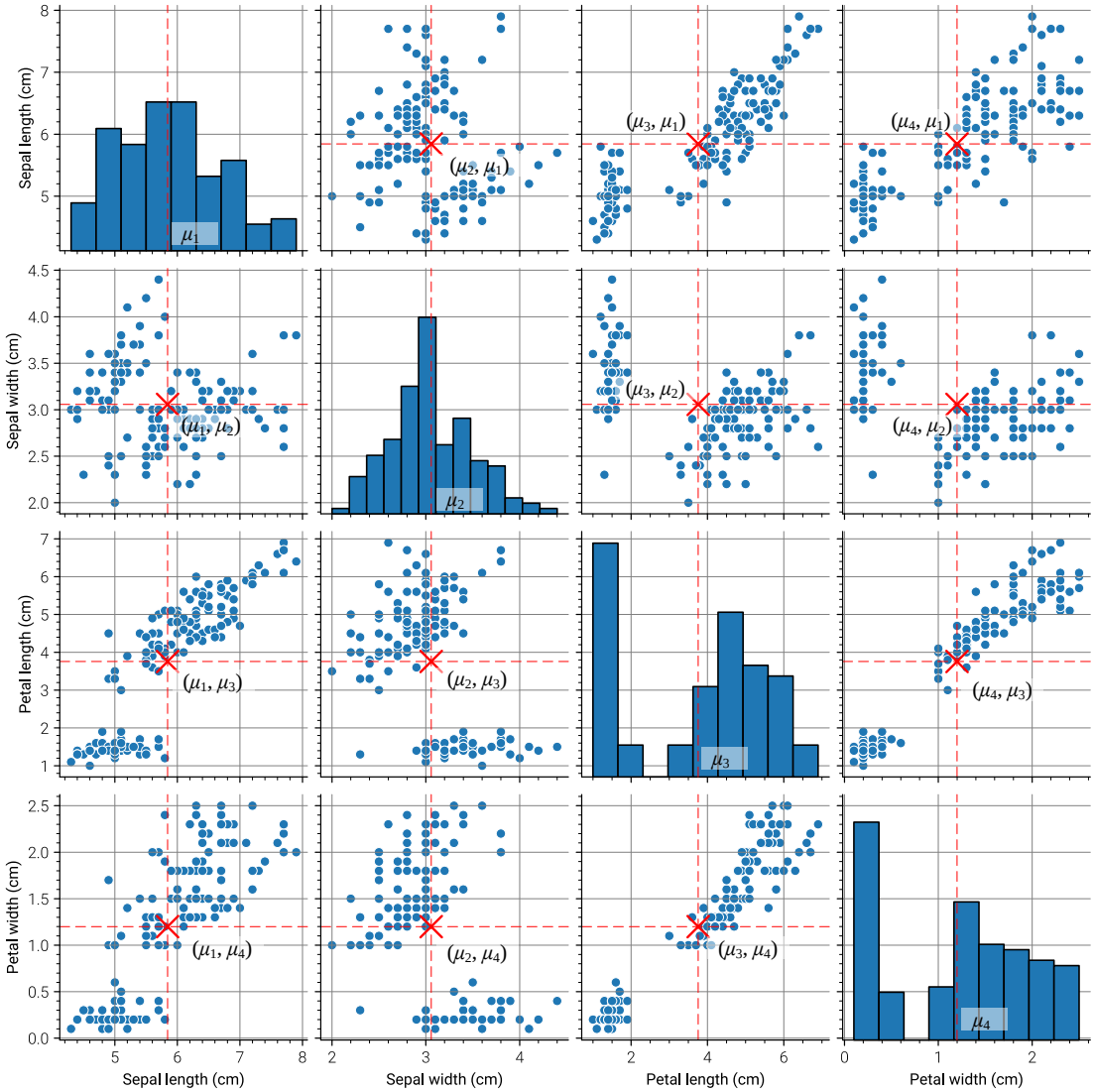

图 6. 鸢尾花数据成对散点图，红 × 为质心位置

数据中心化

数据**中心化** (centralize)，也叫**去均值** (demean)，是指将数据矩阵 $\boldsymbol{X}$ 每列数据减去其所在列的均值，使数据的均值变为零。

数据 (列) 中心化这个过程不会改变数据的分布形状，但会将数据平移到以原点为中心的位置。

几何上来看，就是把数据的质心 $\boldsymbol{\mu}$ 移动到原点 $\boldsymbol{0}$。

比如，数据矩阵 $\boldsymbol{X}$ 的第 1 列减去其均值 μ_1，对应矩阵乘法运算

$$\boldsymbol{x}_1 - \mu_1 \boldsymbol{1} = \boldsymbol{x}_1 - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{x}_1 = \underbrace{\left(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\right)}_{\boldsymbol{M}}\boldsymbol{x}_1 \tag{12}$$

注意，上式中全 1 列向量 $\boldsymbol{1}$ 有 n 行。

令

$$\boldsymbol{M}=\boldsymbol{I}-\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}} \tag{13}$$

我们管 $\boldsymbol{M}$ 叫做中心化矩阵，或去均值矩阵。矩阵 $\boldsymbol{M}$ 为幂等矩阵，即满足 $\boldsymbol{M}@\boldsymbol{M}=\boldsymbol{M}$。

数据矩阵 $\boldsymbol{X}$ 的列中心化，可以通过下式计算得到

$$\boldsymbol{X}_c=\boldsymbol{X}-\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{X}=\underbrace{\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\right)}_{\boldsymbol{M}}\boldsymbol{X}=\boldsymbol{M}\boldsymbol{X} \tag{14}$$

几何上，上式相当于平移。在 (14) “平移”基础上

数据矩阵 $\boldsymbol{X}$ 的行中心化，可以这样算

$$\boldsymbol{X}\underbrace{\left(\boldsymbol{I}-\frac{1}{n}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\right)}_{\boldsymbol{M}}=\boldsymbol{X}\boldsymbol{M} \tag{15}$$

⚠ 注意，(14)、(15) 两式中 $\boldsymbol{M}$ 形状不同。

利用**广播原则** (broadcasting)，(14) 可以写成

$$\boldsymbol{X}_c=\boldsymbol{X}-\boldsymbol{\mu}^{\mathrm{T}} \tag{16}$$

上式相当于每一行行向量完成减法运算。

矩阵运算的广播原则指的是在维度不一致时，自动扩展较小维度的矩阵，使其形状匹配以完成逐元素运算。

图 7 所示为中心化后的鸢尾花数据矩阵热图。

图 8 所示为中心化数据成对散点图。平面散点子图中，我们发现所有质心投影都在原点处。也就是说，质心投影位于零向量 $\boldsymbol{0}$。

对比图 6、图 8 主对角线子图，我们发现数据的分布情况没有任何变化。仅仅是均值位置发生了平移。

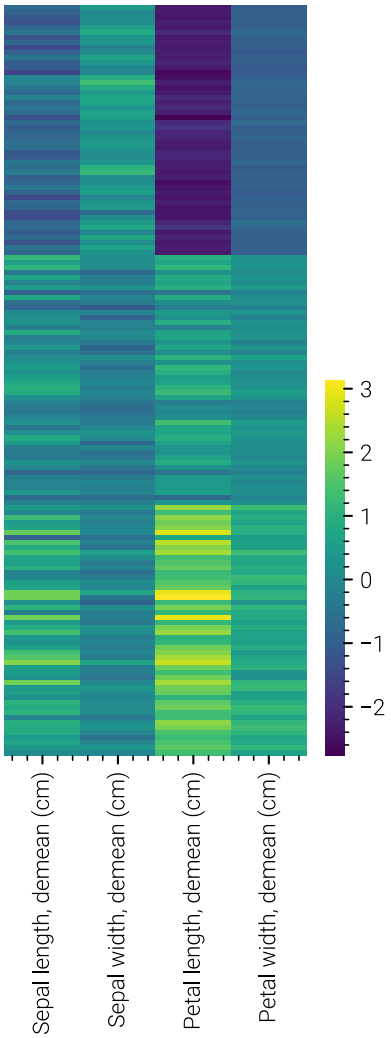

图 7. 鸢尾花数据 (中心化，去均值) 矩阵热图，单位为厘米

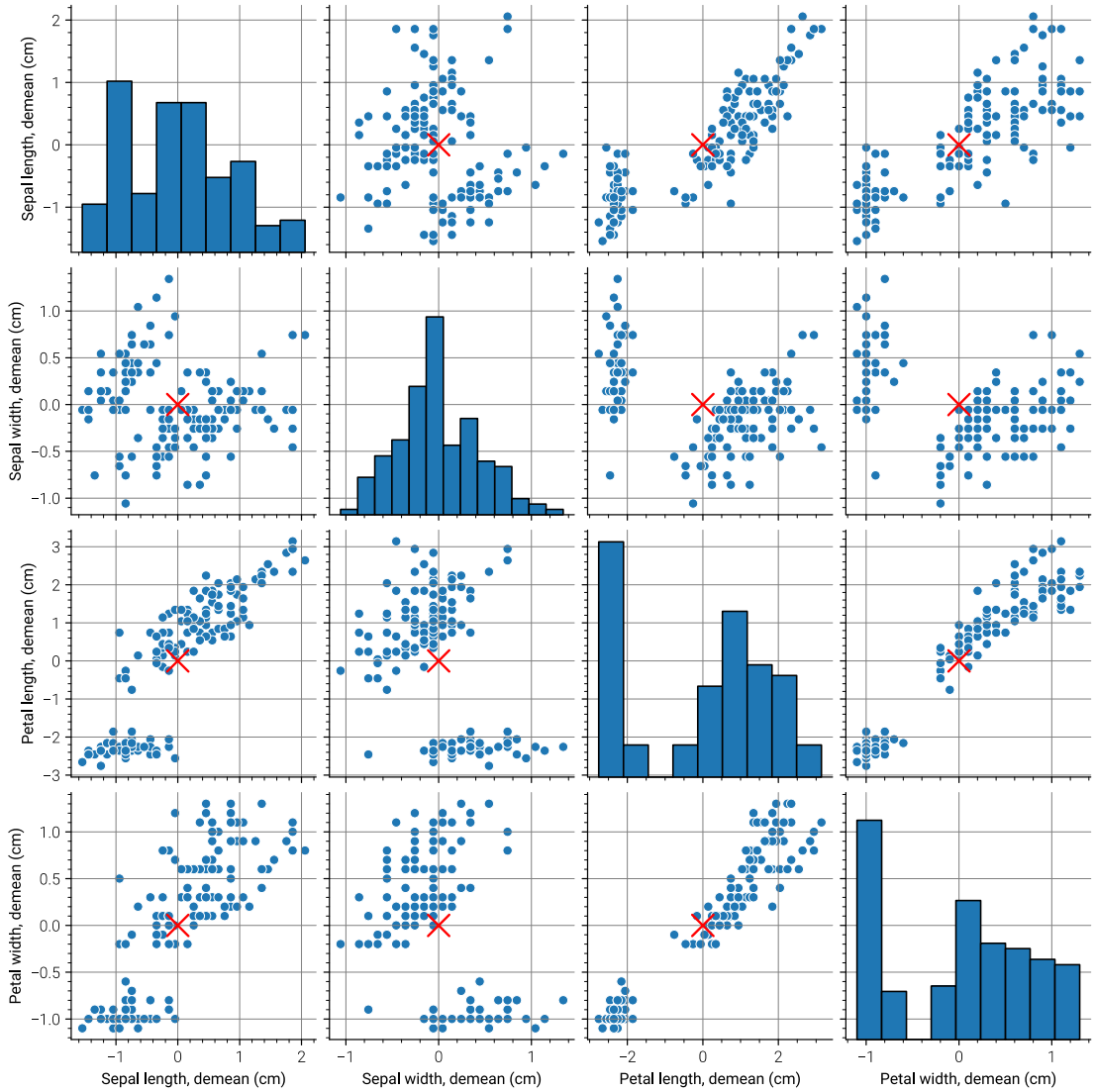

图 8. 鸢尾花数据 (中心化，去均值) 成对散点图，红 × 为质心位置

协方差矩阵

如图 9 所示，有了中心化数据矩阵 $\boldsymbol{X}_c$，我们就可以计算协方差矩阵，对应矩阵乘法

$$\boldsymbol{\Sigma} = \frac{\boldsymbol{X}_c^{\mathrm{T}} \boldsymbol{X}_c}{n-1} \tag{17}$$

计算 $\boldsymbol{X}_c$ 的格拉姆矩阵 $\boldsymbol{X}_c^{\mathrm{T}}\boldsymbol{X}_c$，并用 $1/(n-1)$ 缩放。此外，如果 n 足够大，可以用 n 替换 $n-1$，影响微乎其微。

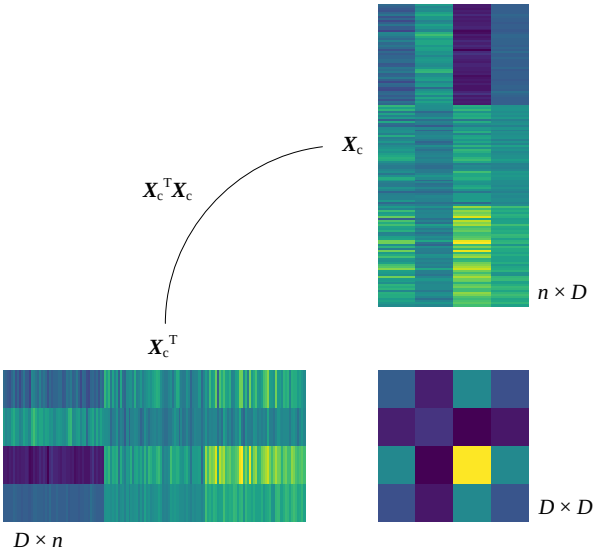

图 9. 计算 $\boldsymbol{X}$ 样本数据协方差矩阵，没有考虑 $n-1$

图 10 所示为鸢尾花数据的协方差矩阵具体数值。

注意，协方差矩阵中方差、协方差的单位为平方厘米。

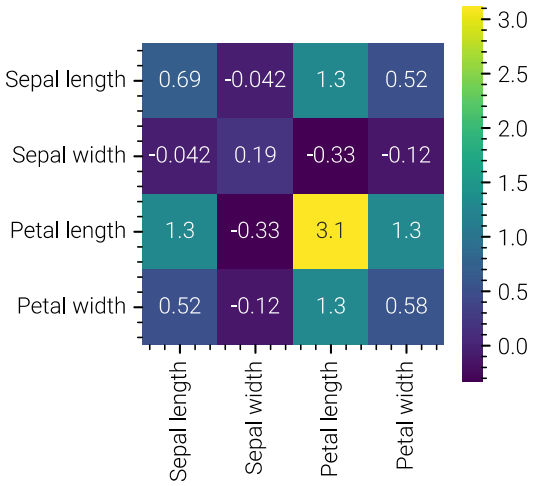

图 10. 协方差矩阵，单位为平方厘米

方差、协方差、线性相关性系数

如图 11 所示，协方差矩阵可视作**方差** (variance) 和**协方差** (covariance) 两部分组成，方差是协方差矩阵对角线上的元素，协方差是协方差矩阵非对角线上的元素：

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D,1} & \sigma_{D,2} & \cdots & \sigma_{D,D} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,D}\sigma_1\sigma_D \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,D}\sigma_2\sigma_D \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D}\sigma_1\sigma_D & \rho_{2,D}\sigma_2\sigma_D & \cdots & \sigma_D^2 \end{bmatrix} \tag{18}$$

方差描述了某个特征上数据的离散度，而协方差则蕴含成对特征之间的线性相关性。上式中，ρ 为**线性相关性系数** (linear correlation coefficient, Pearson correlation coefficient)，常简作**相关性系数**。

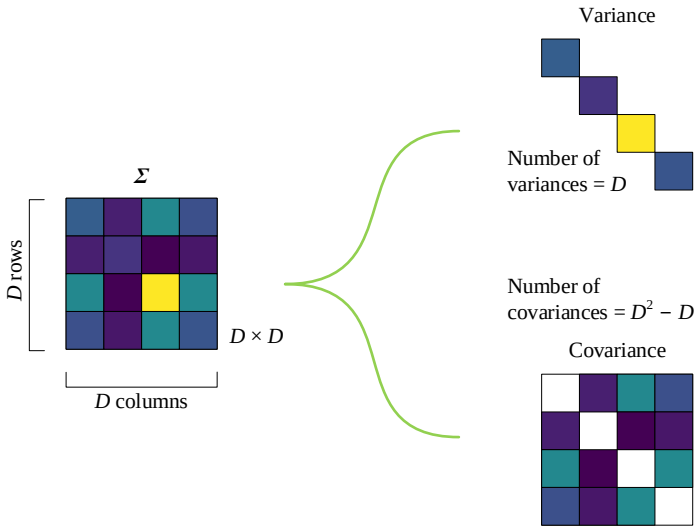

图 11. 协方差矩阵由方差和协方差组成

显而易见，协方差矩阵为对称矩阵：

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{\mathrm{T}} \tag{19}$$

把数据矩阵 $\boldsymbol{X}$ 展开成一组列向量 $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_D]$，(17)可以整理为：

$$\begin{aligned}\boldsymbol{\Sigma} &= \frac{\begin{bmatrix}\boldsymbol{x}_1-\mu_1 & \boldsymbol{x}_2-\mu_2 & \cdots & \boldsymbol{x}_D-\mu_D\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{x}_1-\mu_1 & \boldsymbol{x}_2-\mu_2 & \cdots & \boldsymbol{x}_D-\mu_D\end{bmatrix}}{n-1} \\ &= \frac{1}{n-1}\begin{bmatrix}(\boldsymbol{x}_1-\mu_1)^{\mathrm{T}}(\boldsymbol{x}_1-\mu_1) & (\boldsymbol{x}_1-\mu_1)^{\mathrm{T}}(\boldsymbol{x}_2-\mu_2) & \cdots & (\boldsymbol{x}_1-\mu_1)^{\mathrm{T}}(\boldsymbol{x}_D-\mu_D) \\ (\boldsymbol{x}_2-\mu_2)^{\mathrm{T}}(\boldsymbol{x}_1-\mu_1) & (\boldsymbol{x}_2-\mu_2)^{\mathrm{T}}(\boldsymbol{x}_2-\mu_2) & \cdots & (\boldsymbol{x}_2-\mu_2)^{\mathrm{T}}(\boldsymbol{x}_D-\mu_D) \\ \vdots & \vdots & \ddots & \vdots \\ (\boldsymbol{x}_D-\mu_D)^{\mathrm{T}}(\boldsymbol{x}_1-\mu_1) & (\boldsymbol{x}_D-\mu_D)^{\mathrm{T}}(\boldsymbol{x}_2-\mu_2) & \cdots & (\boldsymbol{x}_D-\mu_D)^{\mathrm{T}}(\boldsymbol{x}_D-\mu_D)\end{bmatrix}\end{aligned} \tag{20}$$

让我们分别看 (20) 中的方差、协方差运算。

比如，下式计算 $\boldsymbol{x}_1$ 的方差

$$\mathrm{var}(\boldsymbol{x}_1) = \sigma_{1,1} = \sigma_1^2 = \frac{(\boldsymbol{x}_1-\mu_1)^{\mathrm{T}}(\boldsymbol{x}_1-\mu_1)}{n-1} \tag{21}$$

大家是否发现上式本质上是向量内积，即

$$\mathrm{var}(\boldsymbol{x}_1) = \sigma_{1,1} = \sigma_1^2 = \frac{(\boldsymbol{x}_1-\mu_1)\cdot(\boldsymbol{x}_1-\mu_1)}{n-1} = \frac{\langle \boldsymbol{x}_1-\mu_1, \boldsymbol{x}_1-\mu_1\rangle}{n-1} \tag{22}$$

方差开平方后得到**标准差** (standard deviation)，即

$$\mathrm{std}(\boldsymbol{x}_1) = \sigma_1 = \sqrt{\mathrm{var}(\boldsymbol{x}_1)} = \sqrt{\frac{(\boldsymbol{x}_1-\mu_1)^{\mathrm{T}}(\boldsymbol{x}_1-\mu_1)}{n-1}} \tag{23}$$

注意，标准差的单位和原始数据、均值一致。比如鸢尾花数据的四个特征的标准差单位均为厘米。

下式计算 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 的协方差

$$\operatorname{cov}(\boldsymbol{x}_1,\boldsymbol{x}_2)=\sigma_{1,2}=\rho_{1,2}\sigma_1\sigma_2=\frac{(\boldsymbol{x}_1-\mu_1)^{\mathrm{T}}(\boldsymbol{x}_2-\mu_2)}{n-1} \tag{24}$$

(24) 也可以写成向量内积，请大家自行完成。

线性相关性系数可以通过下式计算得到

$$\rho_{1,2}=\frac{\operatorname{cov}(\boldsymbol{x}_1,\boldsymbol{x}_2)}{\sigma_1\sigma_2} \tag{25}$$

线性相关性系数的取值范围在 −1 到 1 之间，本质上是对协方差进行归一化处理，使其不再依赖于变量本身的尺度。

协方差虽然能反映两个变量的线性关系方向和强度，但其数值大小受变量单位和尺度影响，不易比较。而相关系数通过除以两个变量的标准差，把协方差标准化到一个固定范围，使得无论变量取值范围如何，都能统一衡量它们之间的线性相关程度。

如图 12 所示，线性相关系数为 1 表示完全正相关，为 −1 表示完全负相关，接近 0 则表示几乎无线性关系。

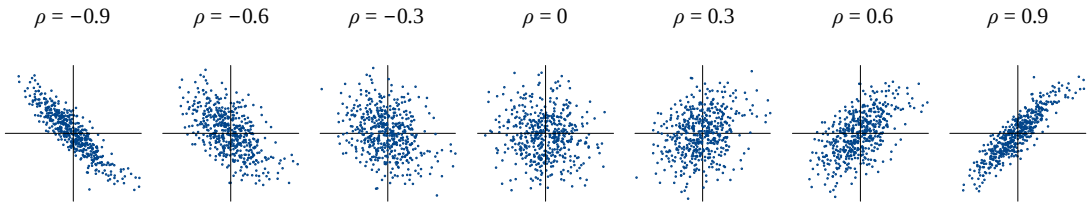

图 12. 随机数散点分布随线性相关性系数变化

质心位于原点

特别地，当所有均值都是 0 时，$[\mu_1, \mu_2, \ldots, \mu_D]^{\mathrm{T}} = [0, 0, \ldots, 0]^{\mathrm{T}}$，也就是说数据质心位于原点，并将 $\boldsymbol{X}$ 写成列向量，(20) 可以写成：

$$\boldsymbol{\Sigma}=\overbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}^{\text{Gram matrix}}=\frac{\boldsymbol{G}}{n-1}=\frac{1}{n-1}\begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_D\\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_D\\ \vdots & \vdots & \ddots & \vdots\\ \boldsymbol{x}_D^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_D^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D^{\mathrm{T}}\boldsymbol{x}_D\end{bmatrix} \tag{26}$$

用向量内积运算，(26) 可以写成：

$$\boldsymbol{\Sigma}=\frac{1}{n-1}\begin{bmatrix}\langle\boldsymbol{x}_1,\boldsymbol{x}_1\rangle & \langle\boldsymbol{x}_1,\boldsymbol{x}_2\rangle & \cdots & \langle\boldsymbol{x}_1,\boldsymbol{x}_D\rangle\\ \langle\boldsymbol{x}_2,\boldsymbol{x}_1\rangle & \langle\boldsymbol{x}_2,\boldsymbol{x}_2\rangle & \cdots & \langle\boldsymbol{x}_2,\boldsymbol{x}_D\rangle\\ \vdots & \vdots & \ddots & \vdots\\ \langle\boldsymbol{x}_D,\boldsymbol{x}_1\rangle & \langle\boldsymbol{x}_D,\boldsymbol{x}_2\rangle & \cdots & \langle\boldsymbol{x}_D,\boldsymbol{x}_D\rangle\end{bmatrix} \tag{27}$$

上式是矩阵乘法的第一视角。

同样，当数据质心位于原点时，将 $\boldsymbol{X}$ 写成行向量，(20) 可以写成：

$$\begin{aligned}\boldsymbol{\Sigma} &= \overbrace{\frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}{n-1}}^{\text{Gram matrix}} = \frac{1}{n-1}\begin{bmatrix}\boldsymbol{x}^{(1)\mathrm{T}} & \boldsymbol{x}^{(2)\mathrm{T}} & \cdots & \boldsymbol{x}^{(n)\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}^{(1)}\\ \boldsymbol{x}^{(2)}\\ \vdots\\ \boldsymbol{x}^{(n)}\end{bmatrix} \\ &= \frac{1}{n-1}\left(\boldsymbol{x}^{(1)\mathrm{T}}\boldsymbol{x}^{(1)} + \boldsymbol{x}^{(2)\mathrm{T}}\boldsymbol{x}^{(2)} + \cdots + \boldsymbol{x}^{(n)\mathrm{T}}\boldsymbol{x}^{(n)}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\boldsymbol{x}^{(i)\mathrm{T}}\boldsymbol{x}^{(i)}\end{aligned} \tag{28}$$

矩阵乘法两个视角

下面用矩阵乘法两个视角来观察 (20)。

根据矩阵乘法第一视角，假设质心为零向量的话，将 $\boldsymbol{X}_{\mathrm{c}}$ 写成 $\begin{bmatrix}\boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D\end{bmatrix}$，(20) 可以展开写成。

$$\mathrm{var}(\boldsymbol{X}) = \boldsymbol{\Sigma} = \frac{1}{n-1}\begin{bmatrix}\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_D \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_D \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_D^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_D^{\mathrm{T}}\boldsymbol{x}_2 & \cdots & \boldsymbol{x}_D^{\mathrm{T}}\boldsymbol{x}_D\end{bmatrix} = \frac{1}{n-1}\begin{bmatrix}\langle \boldsymbol{x}_1, \boldsymbol{x}_1\rangle & \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle & \cdots & \langle \boldsymbol{x}_1, \boldsymbol{x}_D\rangle \\ \langle \boldsymbol{x}_2, \boldsymbol{x}_1\rangle & \langle \boldsymbol{x}_2, \boldsymbol{x}_2\rangle & \cdots & \langle \boldsymbol{x}_2, \boldsymbol{x}_D\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \boldsymbol{x}_D, \boldsymbol{x}_1\rangle & \langle \boldsymbol{x}_D, \boldsymbol{x}_2\rangle & \cdots & \langle \boldsymbol{x}_D, \boldsymbol{x}_D\rangle\end{bmatrix} \tag{29}$$

注意，为了方便写公式，假设上式中 $\boldsymbol{x}_j$ (j = 1, 2, …, D) 已经中心化，即去均值。

如图 13 所示，协方差矩阵的主对角线元素为 $\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_j$，相当于向量内积$\langle \boldsymbol{x}_j, \boldsymbol{x}_j\rangle$，也相当于向量 $\boldsymbol{x}_j$ 的 L2 范数平方 $\|\boldsymbol{x}_j\|_2^2$。

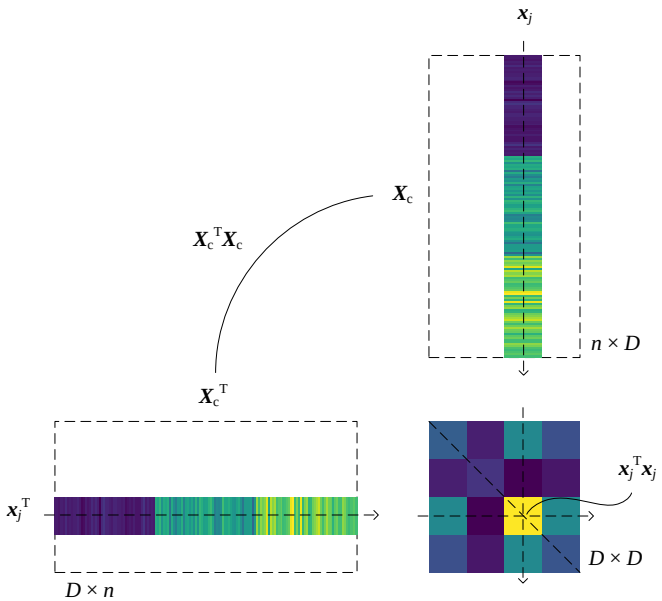

图 13. 协方差矩阵主对角线元素

如图 14 所示，协方差矩阵的非主对角线元素为 $\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_k$ ($j \neq k$)，相当于向量内积$<\boldsymbol{x}_j, \boldsymbol{x}_k>$。显然，$\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_k = \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{x}_k$，即 $<\boldsymbol{x}_j, \boldsymbol{x}_k> = <\boldsymbol{x}_k, \boldsymbol{x}_j>$；这也告诉我们协方差矩阵为对称矩阵。

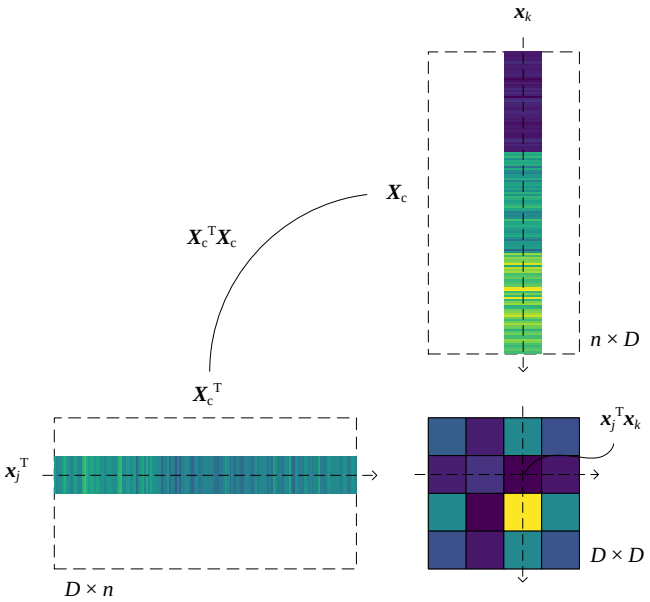

图 14. 协方差矩阵非主对角线元素

正是因为协方差矩阵为对称矩阵，为了减少信息储存量，我们仅仅需要如图 15 所示的这部分矩阵 (方差 + 协方差) 的数据。不管是下三角矩阵还是上三角矩阵，我们保留了 D 个方差、$D(D-1)/2$ 个协方差。也就是，我们保留了 $D(D+1)/2$ 个元素，剔除了 $D(D-1)/2$ 个重复元素。而利用组合数，我们可以容易发现 $C_D^2 = \dfrac{D(D-1)}{2}$，表示在 D 个特征中任意取 2 个特征的组合数。

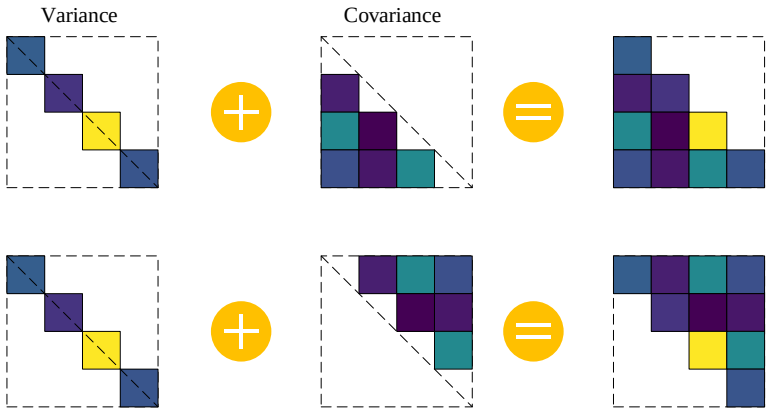

图 15. 剔除协方差矩阵中冗余元素

为了方便讨论，也是假设质心为零向量，根据矩阵乘法第二视角，将 $\boldsymbol{X}_c$ 写成 $\begin{bmatrix} \boldsymbol{x}^{(1)} \\ \boldsymbol{x}^{(2)} \\ \vdots \\ \boldsymbol{x}^{(n)} \end{bmatrix}$，(20) 可以展开写成 n 个秩一矩阵之和，即

$$\boldsymbol{\Sigma} = \frac{1}{n-1}\left[\left(\boldsymbol{x}^{(1)}\right)^{\mathrm{T}}\boldsymbol{x}^{(1)} + \left(\boldsymbol{x}^{(2)}\right)^{\mathrm{T}}\boldsymbol{x}^{(2)} + \cdots + \left(\boldsymbol{x}^{(n)}\right)^{\mathrm{T}}\boldsymbol{x}^{(n)}\right] = \frac{1}{n-1}\sum_{i=1}^{n}\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)} \tag{30}$$

如果 $\boldsymbol{x}^{(i)}$ 不为零向量，每个 $\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)}$ 均为秩一矩阵，形状为 $D \times D$。

如图 16 所示，(30) 相当于对于 n 个 $\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)}$ 取均值；而且，每个样本点都有相同的权重 $\frac{1}{n-1}$。

虽然 $\left(\boldsymbol{x}^{(i)}\right)^{\mathrm{T}}\boldsymbol{x}^{(i)}$ 的秩为 1，但是协方差矩阵 $\boldsymbol{\Sigma}$ 的秩最大为 D，即 $\mathrm{rank}(\boldsymbol{\Sigma}) \leqslant D$。

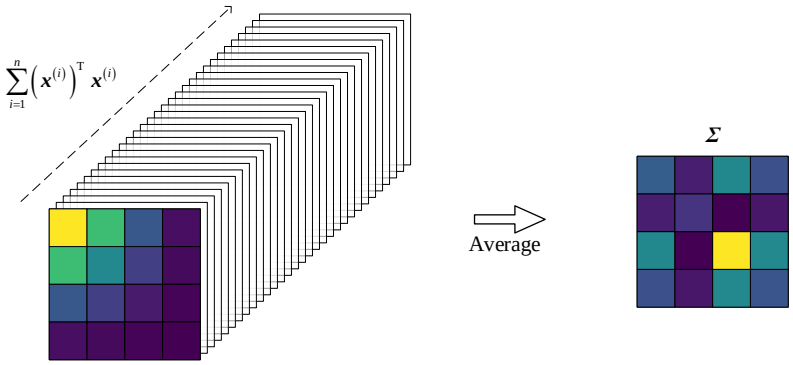

图 16. 协方差矩阵可以看成 n 个秩一矩阵取平均

标准化

数据的**标准化** (standardization) 是将不同尺度的特征转换到相同的尺度，通常是通过减去均值再除以标准差，使得数据的均值为 0、标准差为 1，便于比较不同特征或用于某些对尺度敏感的算法。

令对角方阵 $\boldsymbol{D}$ 为

$$\boldsymbol{D} = \mathrm{diag}\left(\mathrm{diag}\left(\boldsymbol{\Sigma}\right)\right)^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix} \tag{31}$$

里层 diag() 提取协方差主对角线元素，即提取方差，结果为一维数组；外层 diag() 将一维数组构造成对角方阵。

对角方阵的主对角线元素为各个特征的标准差。有了前文的线性代数的基础，大家应该很清楚这个对角方阵的作用相当于缩放。

在 (14) “平移”基础上，再缩放便得到标准化数据矩阵 $\boldsymbol{Z}$，即

$$\boldsymbol{Z} = \boldsymbol{X}_c \boldsymbol{D}^{-1} = \underbrace{\left(\boldsymbol{I} - \frac{1}{n} \boldsymbol{1}\boldsymbol{1}^{\mathrm{T}} \right)}_{\boldsymbol{M}} \boldsymbol{X}\boldsymbol{D}^{-1} = \boldsymbol{M}\boldsymbol{X}\boldsymbol{D}^{-1} \tag{32}$$

其中，

$$\boldsymbol{D}^{-1} = \operatorname{diag}\left(\operatorname{diag}\left(\boldsymbol{\Sigma}\right)\right)^{\frac{-1}{2}} = \begin{bmatrix} 1/\sigma_1 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_D \end{bmatrix} \tag{33}$$

几何上，(32) 相当于完成“平移 → 缩放”。

图 17 所示为鸢尾花标准化数据成对散点图。

标准化后的数据被转换为**无单位** (unitless) 的形式，是因为标准化操作将每个数值减去其特征的均值后，再除以标准差。

这样处理后的结果表示的是“距离均值多少个标准差”，而不是原始的物理量，因此不再携带原来的单位。

例如，将花萼长度从厘米标准化后，结果就表示该鸢尾花样本的花萼长度比平均长度大或小多少个标准差，而不再是“多少厘米”。这使得不同单位、不同量纲的数据可以在同一尺度上进行比较和建模

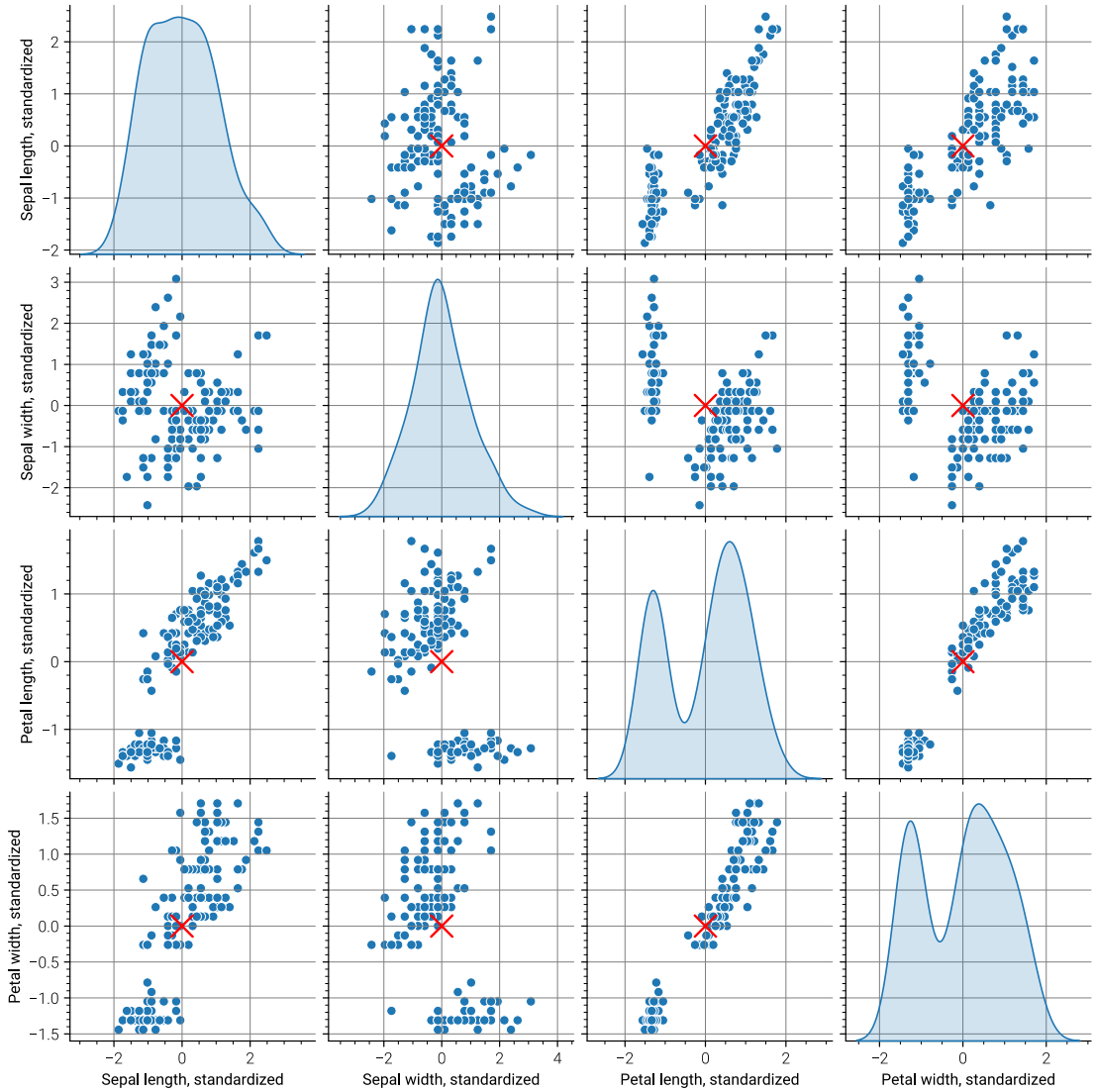

图 17. 鸢尾花数据 (标准化) 成对散点图，红 × 为质心位置，没有单位

线性相关性系数矩阵

线性相关性系数矩阵 $\boldsymbol{P}$ 的定义为：

$$\boldsymbol{P}=\begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix} \tag{34}$$

图 18 所示为鸢尾花数据相关性系数矩阵 $\boldsymbol{P}$。$\boldsymbol{P}$ 的对角线元素均为 1，对角线以外元素为成对相关性系数 $\rho_{i,j}$。

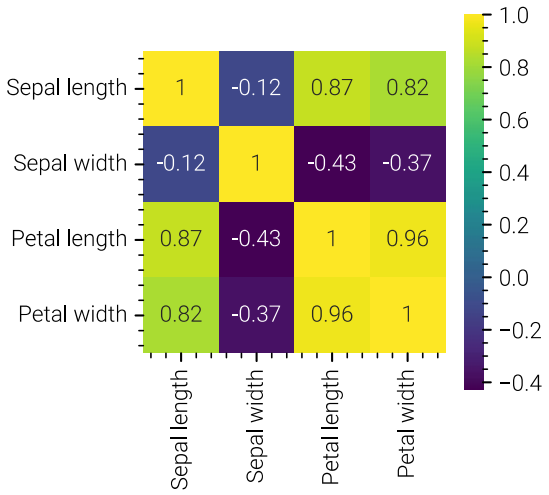

图 18. 线性相关性系数矩阵

协方差矩阵 $\boldsymbol{\Sigma}$ 和相关性系数矩阵 $\boldsymbol{P}$ 关系如下：

$$\boldsymbol{\Sigma} = \boldsymbol{DPD} = \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix}}_{\boldsymbol{D}} \underbrace{\begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,D} \\ \rho_{1,2} & 1 & \cdots & \rho_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D} & \rho_{2,D} & \cdots & 1 \end{bmatrix}}_{\text{Correlation matrix, } \boldsymbol{P}} \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_D \end{bmatrix}}_{\boldsymbol{D}} \tag{35}$$

从几何角度来看，上式中对角方阵 $\boldsymbol{D}$ 起到的是缩放作用。

图 19 所示为协方差矩阵和相关性矩阵关系热图。

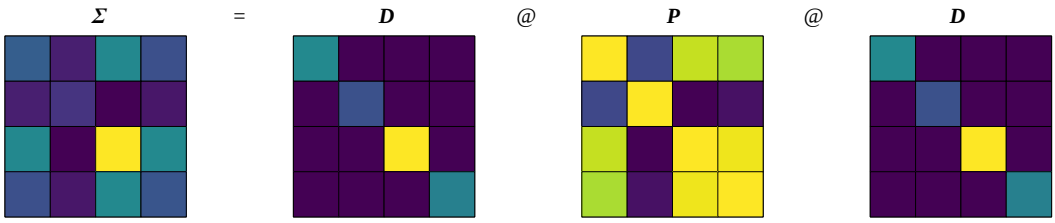

图 19. 协方差矩阵和相关性矩阵关系热图

从 $\boldsymbol{\Sigma}$ 反求相关性系数矩阵 $\boldsymbol{P}$

$$\boldsymbol{P} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1} \tag{36}$$

相信细心的读者可能已经发现，线性相关性系数矩阵 $\boldsymbol{P}$ 相当于标准化数据矩阵的协方差矩阵，即

$$\boldsymbol{P} = \boldsymbol{\Sigma}_z = \overbrace{\frac{\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}}{n-1}}^{\text{Gram matrix}} = \frac{1}{n-1}\begin{bmatrix} \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_1 & \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_2 & \cdots & \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{z}_D \\ \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_1 & \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_2 & \cdots & \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{z}_D \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{z}_D^{\mathrm{T}}\boldsymbol{z}_1 & \boldsymbol{z}_D^{\mathrm{T}}\boldsymbol{z}_2 & \cdots & \boldsymbol{z}_D^{\mathrm{T}}\boldsymbol{z}_D \end{bmatrix} \tag{37}$$

如图 20 所示，上式也相当于计算格拉姆矩阵。

请大家分析图 20 的主对角线、非主对角线元素对应的运算，以及统计意义。

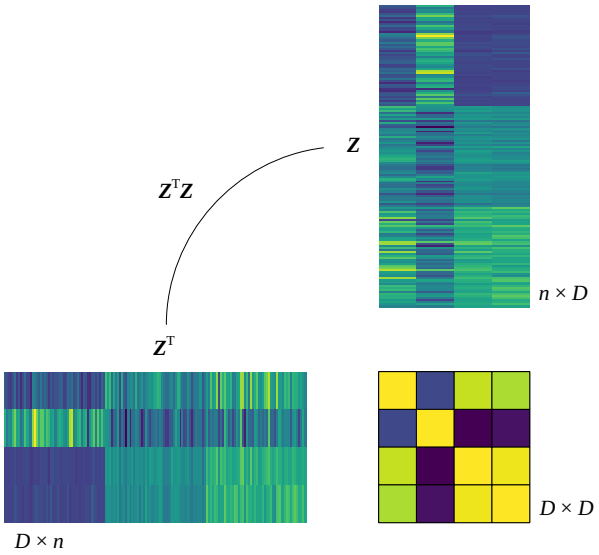

图 20. 计算 $\boldsymbol{Z}$ 标准化数据协方差矩阵 ($\boldsymbol{X}$ 的线性相关性系数矩阵)，没有考虑 $n-1$

LA_12_01_01.ipynb 完成本节所有运算以及可视化，请大家自学。

请大家用 DeepSeek/ChatGPT 等工具完成本节如下习题。

Q1. 请修改 LA_12_01_01.ipynb，对协方差矩阵特征值分解，并且分析特征值、特征向量矩阵的各种性质。

Q2. 请修改 LA_12_01_01.ipynb，对线性相关性系数矩阵特征值分解，并且分析特征值、特征向量矩阵的各种性质。

Q3. 请大家自学广播原则：

https://numpy.org/doc/stable/user/basics.broadcasting.html