

5

Distance Measures in Data

数据距离

距离不仅仅是两点之间的直线线段



当一匹马需要赶超马群时，它才能超越自己。

A horse never runs so fast as when he has other horses to catch up and outpace.

—— 奥维德 (Ovid) | 古罗马诗人 | 43 BC ~ 17/18 AD



```

< scipy.spatial.distance.chebyshev() 计算切比雪夫距离
< scipy.spatial.distance.cityblock() 计算城市街区距离
< scipy.spatial.distance.euclidean() 计算欧氏距离
< scipy.spatial.distance.mahalanobis() 计算马氏距离
< scipy.spatial.distance.minkowski() 计算闵氏距离
< scipy.spatial.distance.seuclidean() 计算标准化欧氏距离
< seaborn.scatterplot() 绘制散点图
< sklearn.datasets.load_iris() 加载鸢尾花数据集
< sklearn.metrics.pairwise.euclidean_distances() 计算成对欧氏距离矩阵
< sklearn.metrics.pairwise_distances() 计算成对距离矩阵
< metrics.pairwise.linear_kernel() 计算线性核成对亲密度矩阵
< metrics.pairwise.manhattan_distances() 计算成对城市街区距离矩阵
< metrics.pairwise.paired_cosine_distances(X,Q) 计算 X 和 Q 样本数据矩阵成对余弦距离矩阵
< metrics.pairwise.paired_euclidean_distances(X,Q) 计算 X 和 Q 样本数据矩阵成对欧氏距离矩阵
< metrics.pairwise.paired_manhattan_distances(X,Q) 计算 X 和 Q 样本数据矩阵成对城市街区距离矩阵
< metrics.pairwise.polynomial_kernel() 计算多项式核成对亲密度矩阵
< metrics.pairwise.rbf_kernel() 计算 RBF 核成对亲密度矩阵
< metrics.pairwise.sigmoid_kernel() 计算 sigmoid 核成对亲密度矩阵

```

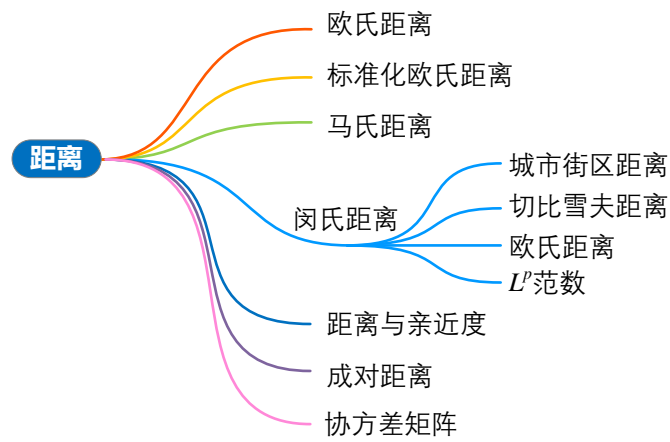
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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



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

5.1 怎么又聊距离？

“鸢尾花书”似乎对距离特别“痴迷”，几乎每个分册都会聊到距离相关内容。一方面是因为距离这个概念本身的外延很广，很多数学工具都可以从距离这个几何视角来观察；此外，机器学习中大部分算法都离不开距离。

距离在机器学习中发挥着重要作用，通常用于衡量数据点之间的相似性或差异性。下面让我们一起举几个例子聊聊数据分析、机器学习中的距离。

如图 1 所示，从几何角度来看，一元线性回归就是在 \mathbf{I} (全 1 列向量) 和 \mathbf{x} 构成平面内找到 \mathbf{y} 的投影，使得 ϵ 尽可能小。 ϵ 本质上就是距离。

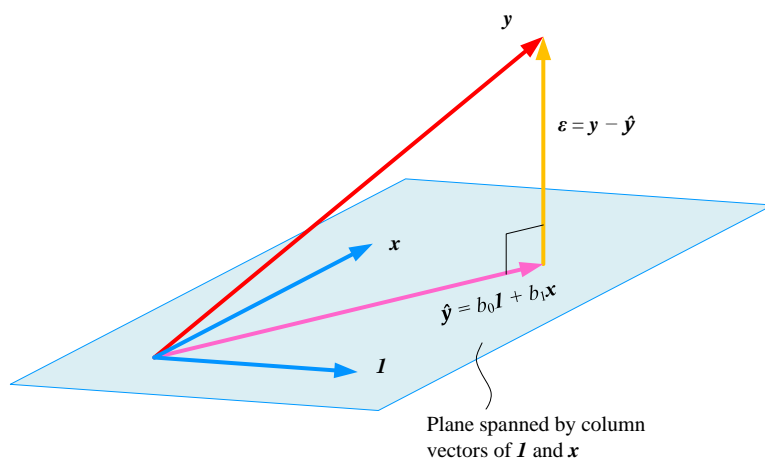


图 1. 几何角度解释一元线性回归最小二乘结果

如图 2 所示，**主成分分析** (Principal Component Analysis, PCA) 中，选取第一主成分 \mathbf{v} 的标准是—— z 方差最大化。方差可以看做一种距离，标准差也是距离；连 Z 分数也可以看成是一种距离。从这个角度来看，协方差矩阵就是距离的集合体。

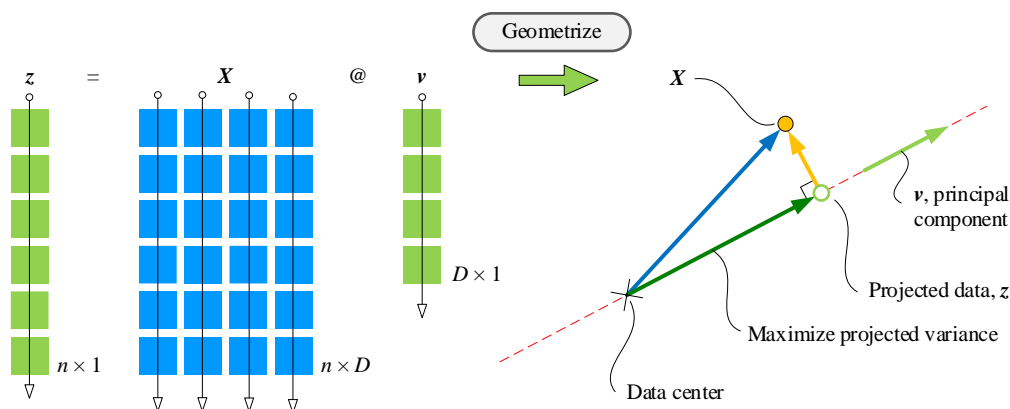


图 2. 主成分分析优化问题

如图 3 所示，**支持向量机** (Support Vector Machine, SVM) 算法中，我们则关心支持向量到决策平面的距离。

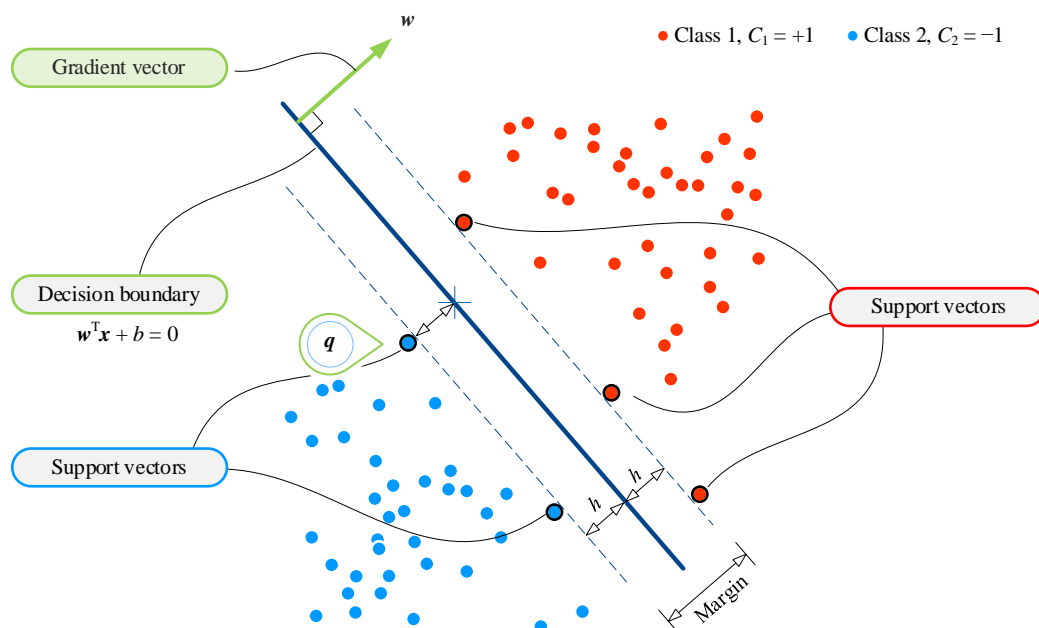


图 3. 支持向量机

如图 4 所示，**层次聚类** (hierarchical clustering) 中，我们不但关注数据点之间的距离，还需要计算簇间距离。

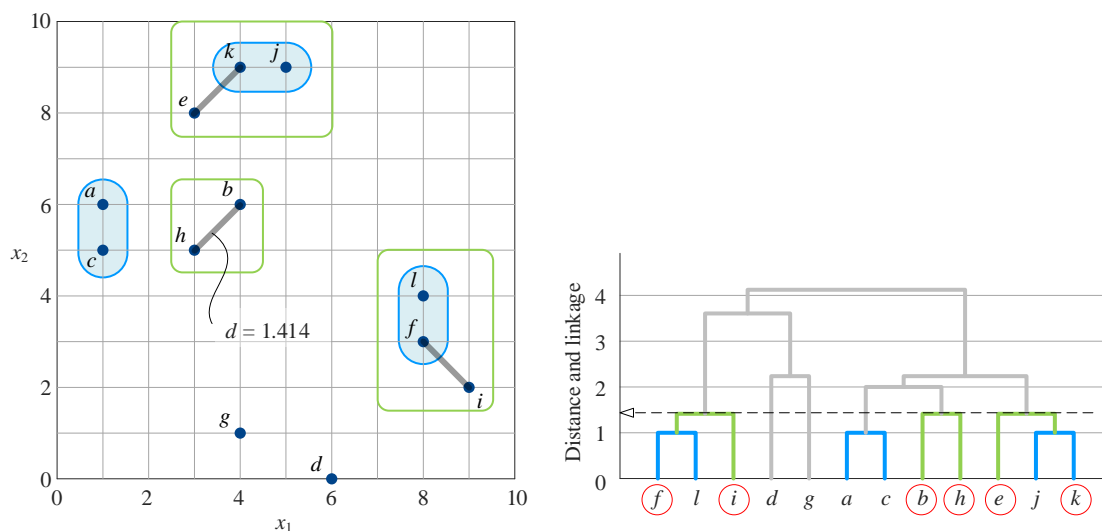


图 4. 层次聚类中构建树形图，第二层



大家对距离这个概念应该非常熟悉，我们从《数学要素》第 7 章开始就不断丰富“距离”的内涵。我们在《矩阵力量》第 3 章专门介绍了基于 L^p 范数的几种距离度量，在《统计至简》第 15 章专门讲解了马氏距离。

本章后续专门总结并探讨常用的几个距离度量。

- ◀ 欧氏距离 (Euclidean distance)
- ◀ 标准化欧氏距离 (standardized Euclidean distance)
- ◀ 马氏距离 (Mahalanobis distance, Mahal distance)
- ◀ 城市街区距离 (city block distance)
- ◀ 切比雪夫距离 (Chebyshev distance)
- ◀ 闵氏距离 (Minkowski distance)
- ◀ 余弦距离 (cosine distance)
- ◀ 相关性距离 (correlation distance)

本章最后将在距离的视角下再看协方差矩阵。

5.2 欧氏距离：最常见的距离

欧几里得距离，也称**欧氏距离** (Euclidean distance)。欧氏距离是机器学习中常用的一种距离度量方法，适用于处理连续特征的数据。其特点是简单易懂、计算效率高，但容易受到数据维度、特征尺度、特征量纲影响。

任意样本数据点 \mathbf{x} 和查询点 \mathbf{q} 欧氏距离定义如下：

$$d(\mathbf{x}, \mathbf{q}) = \|\mathbf{x} - \mathbf{q}\| = \sqrt{(\mathbf{x} - \mathbf{q})^T (\mathbf{x} - \mathbf{q})} \quad (1)$$

其中， \mathbf{x} 和 \mathbf{q} 为列向量。欧氏距离本质上就是 $\mathbf{x} - \mathbf{q}$ 的 L^2 范数。从几何视角来看，二维欧氏距离可以看做同心正圆，三维欧氏距离可以视作同心正球体，等等。

当特征数为 D 时，上式展开可以得到：

$$d(\mathbf{x}, \mathbf{q}) = \sqrt{(x_1 - q_1)^2 + (x_2 - q_2)^2 + \dots + (x_D - q_D)^2} \quad (2)$$

特别地，当特征数量 $D = 2$ 时， \mathbf{x} 和 \mathbf{q} 两点欧氏距离定义为：

$$d(\mathbf{x}, \mathbf{q}) = \sqrt{(x_1 - q_1)^2 + (x_2 - q_2)^2} \quad (3)$$

举个例子

如果查询点 \mathbf{q} 有两个特征，并位于原点，即：

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

$$\mathbf{q} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (4)$$

如图 5 所示，三个样本点 $\mathbf{x}^{(1)}$ 、 $\mathbf{x}^{(2)}$ 和 $\mathbf{x}^{(3)}$ 的位置如下：

$$\mathbf{x}^{(1)} = [-5 \ 0], \ \mathbf{x}^{(2)} = [4 \ 3], \ \mathbf{x}^{(3)} = [3 \ -4] \quad (5)$$

根据 (1) 可以计算得到三个样本点 $\mathbf{x}^{(1)}$ 、 $\mathbf{x}^{(2)}$ 和 $\mathbf{x}^{(3)}$ 距离查询点 \mathbf{q} 之间欧氏距离均为 5：

$$\begin{cases} d_1 = \sqrt{([0 \ 0] - [-5 \ 0])([0 \ 0] - [-5 \ 0])^T} = \sqrt{[5 \ 0][5 \ 0]^T} = \sqrt{25+0} = 5 \\ d_2 = \sqrt{([0 \ 0] - [4 \ 3])([0 \ 0] - [4 \ 3])^T} = \sqrt{[-4 \ -3][-4 \ -3]^T} = \sqrt{16+9} = 5 \\ d_3 = \sqrt{([0 \ 0] - [3 \ -4])([0 \ 0] - [3 \ -4])^T} = \sqrt{[-3 \ 4][-3 \ 4]^T} = \sqrt{9+16} = 5 \end{cases} \quad (6)$$

⚠ 注意，行向量和列向量的转置关系，本章后续不再区分行、列向量。

如图 5 所示，当 d 取定值时，上式相当于以 (q_1, q_2) 为圆心的正圆。

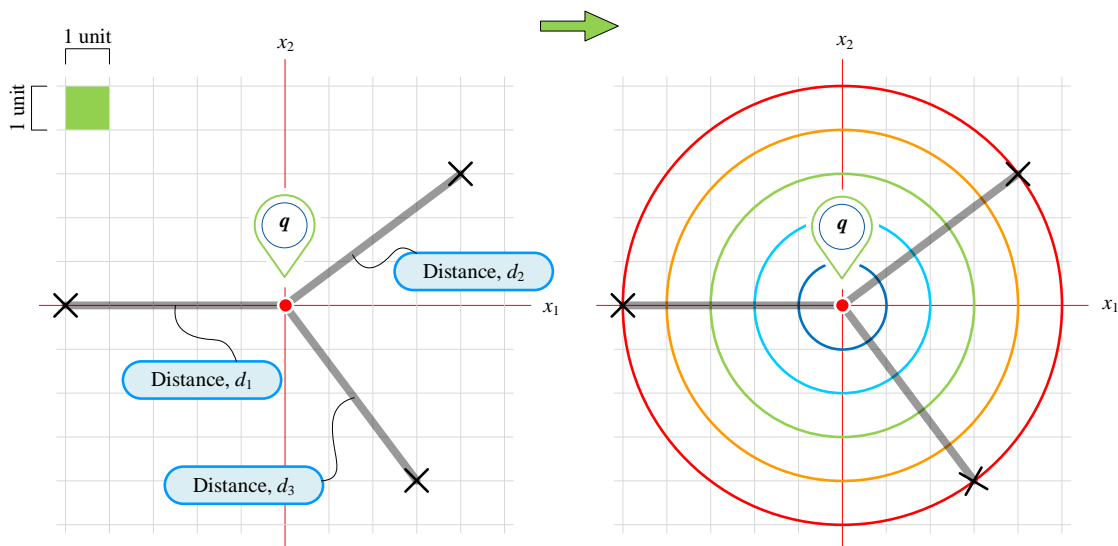


图 5.2 特征 ($D=2$) 欧几里得距离



代码 Bk6_Ch05_01.ipynb 计算两点欧氏距离。`scipy.spatial.distance.euclidean()` 为计算欧氏距离的函数。

成对距离

如图 5 所示，三个样本点 $\mathbf{x}^{(1)}$ 、 $\mathbf{x}^{(2)}$ 和 $\mathbf{x}^{(3)}$ 之间也存在两两距离，我们管它们叫做**成对距离** (pairwise distance)。图 6 所示为平面上 12 个点的成对距离。成对距离结果一般以矩阵方式呈现。

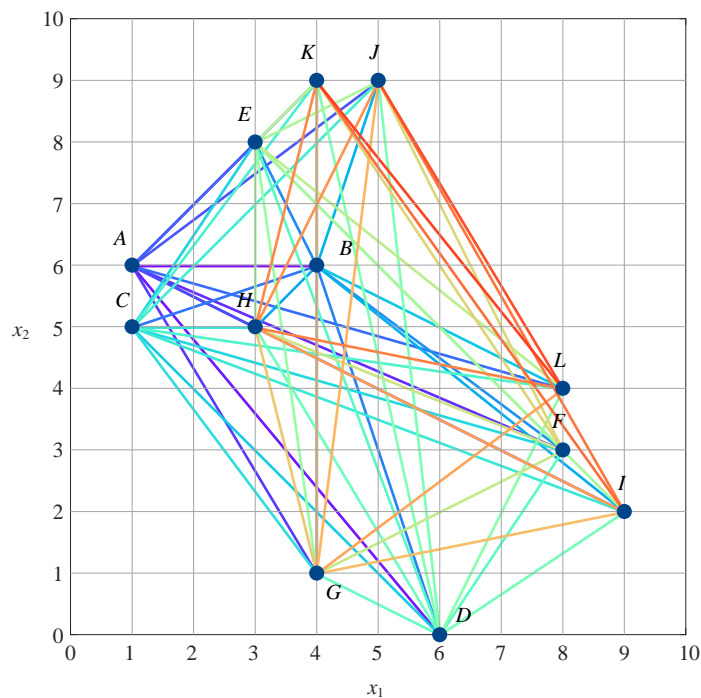
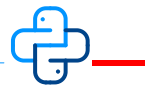


图 6. 平面上 12 个点，成对距离，来自鸢尾花书《数学要素》



代码 Bk6_Ch05_02.ipynb 计算图 5 中三个样本点之间的成对欧氏距离。

5.3 标准化欧氏距离：考虑标准差

标准化欧氏距离 (standardized Euclidean distance) 是一种将欧氏距离进行归一化处理的方法，适用于处理特征间尺度差异较大的数据。其特点是能够消除不同特征之间的度量单位和尺度差异，从而减少距离计算结果偏差。优点是比欧氏距离更具有鲁棒性和稳定性，缺点是对于一些特征较为稀疏的数据，可能存在一些计算上的困难。

定义

标准化欧氏距离定义如下。

$$d(\mathbf{x}, \mathbf{q}) = \sqrt{(\mathbf{x} - \mathbf{q})^T \mathbf{D}^{-1} \mathbf{D}^{-1} (\mathbf{x} - \mathbf{q})} \quad (7)$$

其中， \mathbf{D} 为对角方阵，对角线元素为标准差，运算如下：

$$D = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}} = \text{diag} \left(\text{diag} \left[\begin{array}{cccc} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,D}\sigma_1\sigma_D \\ \rho_{1,2}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & \rho_{2,D}\sigma_2\sigma_D \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,D}\sigma_1\sigma_D & \rho_{2,D}\sigma_2\sigma_D & \cdots & \sigma_D^2 \end{array} \right] \right)^{\frac{1}{2}} = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_D \end{bmatrix} \quad (8)$$

回忆《矩阵力量》介绍过有关 `diag()` 函数的说明。如果 A 为方阵，`diag(A)` 函数提取对角线元素，结果为向量；如果 a 为向量，`diag(a)` 函数将向量 a 展开成对角方阵，方阵对角线元素为 a 向量元素。NumPy 中完成这一计算的函数为 `numpy.diag()`。

将 (8) 带入 (7) 得到：

$$d(\mathbf{x}, \mathbf{q}) = \sqrt{[\mathbf{x} - \mathbf{q}] \begin{bmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_D^2 \end{bmatrix}^{-1} [\mathbf{x} - \mathbf{q}]^T} \quad (9)$$

$$= \sqrt{\frac{(x_1 - q_1)^2}{\sigma_1^2} + \frac{(x_2 - q_2)^2}{\sigma_2^2} + \cdots + \frac{(x_D - q_D)^2}{\sigma_D^2}} = \sqrt{\sum_{j=1}^D \left(\frac{x_j - q_j}{\sigma_j} \right)^2}$$

(9) 可以记做：

$$d(\mathbf{x}, \mathbf{q}) = \sqrt{z_1^2 + z_2^2 + \cdots + z_D^2} = \sqrt{\sum_{j=1}^D z_j^2} \quad (10)$$

其中， z_j 为：

$$z_j = \frac{x_j - q_j}{\sigma_j} \quad (11)$$

上式本质上就是 Z 分数。



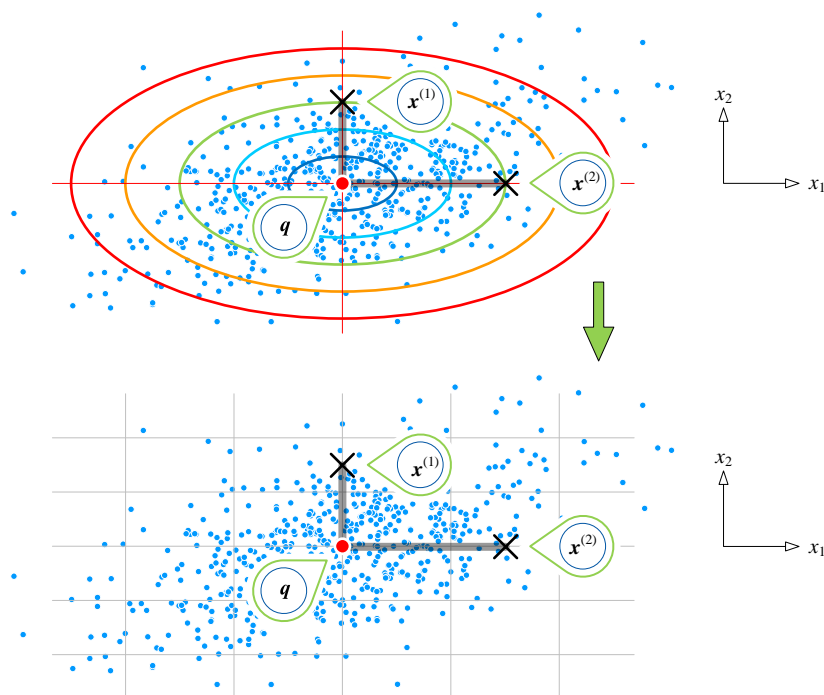
《统计至简》第 9 章专门介绍 Z 分数，请大家回顾。

正椭圆

对于 $D = 2$ ，两特征的情况，标准化欧氏距离平方可以写成：

$$d^2 = \frac{(x_1 - q_1)^2}{\sigma_1^2} + \frac{(x_2 - q_2)^2}{\sigma_2^2} \quad (12)$$

可以发现，上式代表的形状是以 (q_1, q_2) 为中心的正椭圆。观察 (12)，可以发现，标准化欧氏距离引入数据每个特征标准差，但是没有考虑特征之间的相关性。图 7 中，网格的坐标已经转化为“标准差”，而标准欧氏距离等距线为正椭圆。

图 7.2 特征 ($D = 2$) 标准化欧氏距离

几何变换视角

如图 8 所示，从几何变换角度，标准化欧氏距离相当于对 \mathbf{X} 数据每个维度，首先**中心化** (centralize)，然后利用标准差进行**缩放** (scale)；但是，标准化欧氏距离没有旋转操作，也就是没有正交化。

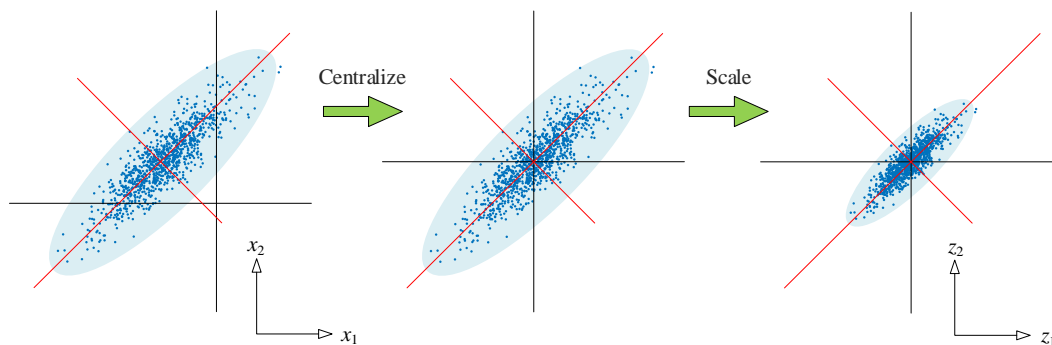


图 8. 标准化欧氏距离运算过程



计算标准化欧氏距离的函数为 `scipy.spatial.distance.seuclidean()`。代码 Bk6_Ch05_03.ipynb 计算本节标准化欧氏距离。

5.4 马氏距离：考虑标准差和相关性



本系列丛书《矩阵力量》和《统计至简》从不同角度讲过马氏距离，本节稍作回忆。

马氏距离 (Mahalanobis distance, Mahal distance)，又叫**马哈距离**，全称马哈拉诺比斯距离，是机器学习中常用的一种距离度量方法，适用于处理高维数据和特征之间存在相关性的情况。其特点是考虑到特征之间的相关性，从而在计算距离时可以更好地描述数据之间的相似程度。优点是能够提高模型的准确性，缺点是对于样本数较少的情况下容易过拟合，计算量较大，同时对数据的分布形式存在假设前提（多元正态分布）。

马氏距离定义如下：

$$d(\mathbf{x}, \mathbf{q}) = \sqrt{(\mathbf{x} - \mathbf{q})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{q})} \quad (13)$$

其中， $\boldsymbol{\Sigma}$ 为协方差矩阵， \mathbf{q} 一般是样本数据的质心。



注意，马氏距离的单位是“标准差”。比如，马氏距离计算结果为 3，应该称作 3 个标准差。

特征值分解：缩放 → 旋转 → 平移

$\boldsymbol{\Sigma}$ 谱分解得到：

$$\boldsymbol{\Sigma} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T \quad (14)$$

其中， \mathbf{V} 为正交矩阵。

$\boldsymbol{\Sigma}^{-1}$ 的特征值分解可以写成：

$$\boldsymbol{\Sigma}^{-1} = \mathbf{V} \mathbf{\Lambda}^{-1} \mathbf{V}^T \quad (15)$$

将 (15) 代入 (13) 得到：

$$d(\mathbf{x}, \boldsymbol{\mu}) = \left\| \begin{matrix} \frac{-1}{\Lambda^2} & \mathbf{V}^T \\ \text{Scale} & \text{Rotate} & \text{Centralize} \end{matrix} \begin{pmatrix} \mathbf{x} - \boldsymbol{\mu} \end{pmatrix} \right\| \quad (16)$$

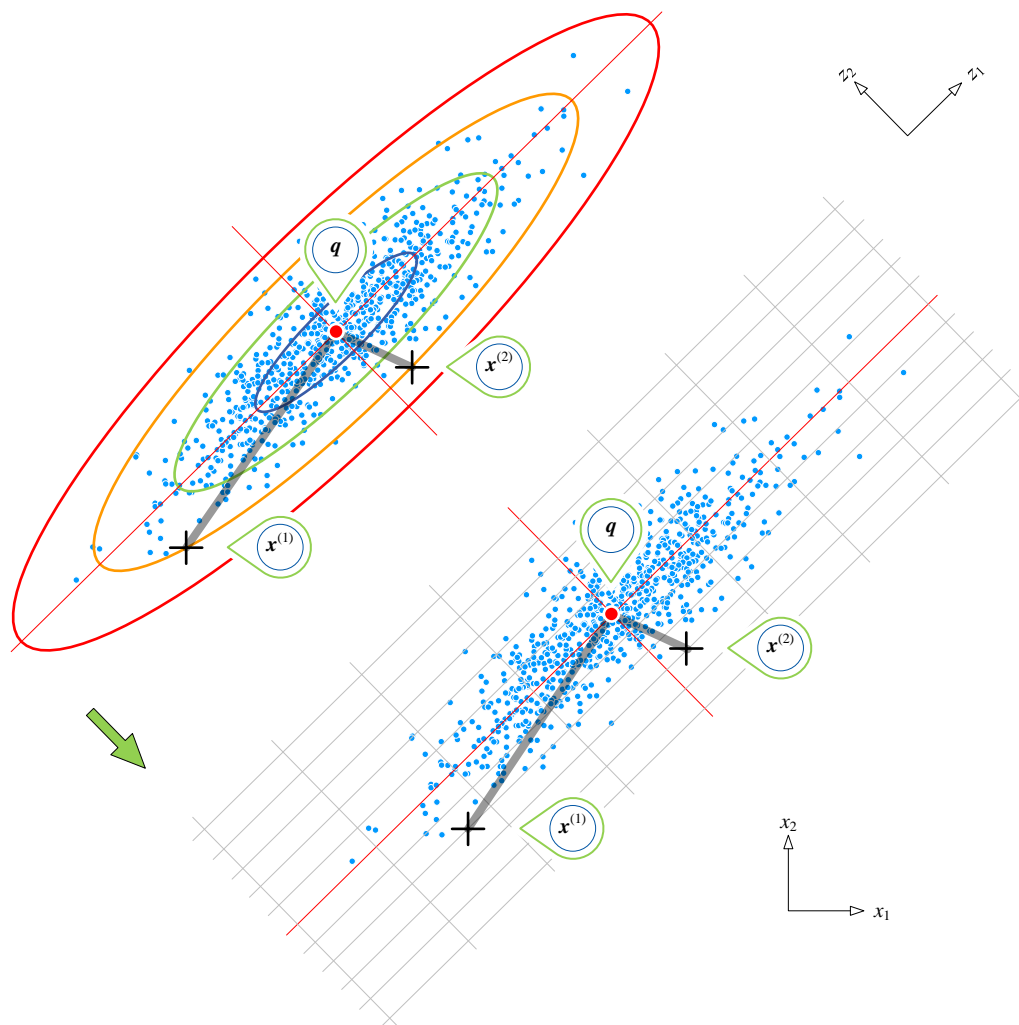
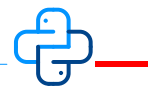
其中， $\boldsymbol{\mu}$ 列向量完成**中心化** (centralize)， \mathbf{V} 矩阵完成**旋转** (rotate)， $\mathbf{\Lambda}$ 矩阵完成**缩放** (scale)。

旋转椭圆

如图 9 所示，当 $D = 2$ 时，马氏距离的等距线为旋转椭圆。



大家如果对这部分内容感到陌生，请回顾《矩阵力量》第 20 章、《统计至简》第 23 章。

图 9.2 特征 ($D=2$) 马氏距离

代码 Bk6_Ch05_04.ipynb 计算图 9 两个点的马氏距离。

举例

下面，我们用具体数字举例讲解如何计算马氏距离。

给定质心 $\mu = [0, 0]^T$ 。两个样本点的坐标分别为。

$$\mathbf{x}^{(1)} = [-3.5 \quad -4]^T, \quad \mathbf{x}^{(2)} = [2.75 \quad -1.5]^T \quad (17)$$

计算得到 $\mathbf{x}^{(1)}$ 和 $\mathbf{x}^{(2)}$ 距离 μ 之间欧氏距离 (L^2 范数) 分别为 5.32 和 3.13。

假设方差协方差矩阵 Σ 取值如下。

$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad (18)$$

观察如上矩阵，可以发现 x_1 和 x_2 特征各自的方差均为 2，两者协方差为 1；计算得到 x_1 和 x_2 特征相关性为 0.5。根据 Σ 计算 $\mathbf{x}^{(1)}$ 和 $\mathbf{x}^{(2)}$ 距离 μ 之间马氏距离为。

$$\begin{aligned} d_1 &= \sqrt{([[-3.5 \quad -4] - [0 \quad 0]] \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} ([[-3.5 \quad -4] - [0 \quad 0]])^T} \\ &= \sqrt{[-3.5 \quad -4] \cdot \frac{1}{3} \cdot \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} [-3.5 \quad -4]^T} = 3.08 \\ d_2 &= \sqrt{([[-2.75 \quad -1.5] - [0 \quad 0]] \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}^{-1} ([[-2.75 \quad -1.5] - [0 \quad 0]])^T} \\ &= \sqrt{[-2.75 \quad -1.5] \cdot \frac{1}{3} \cdot \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} [-2.75 \quad -1.5]^T} = 3.05 \end{aligned} \quad (19)$$

可以发现， $\mathbf{x}^{(1)}$ 和 $\mathbf{x}^{(2)}$ 和 μ 之间马氏距离非常接近。

5.5 城市街区距离： L^1 范数

城市街区距离 (city block distance)，也称**曼哈顿距离** (Manhattan distance)，和欧氏距离本质上都是 L^p 范数。请大家注意区别两者等高线。

城市街区距离具体定义如下：

$$d(\mathbf{x}, \mathbf{q}) = \|\mathbf{x} - \mathbf{q}\|_1 = \sum_{j=1}^D |x_j - q_j| \quad (20)$$

其中， j 代表特征序号。



城市街区距离就是我们在《矩阵力量》第 3 章中介绍的 L^1 范数。

将 (20) 展开得到下式：

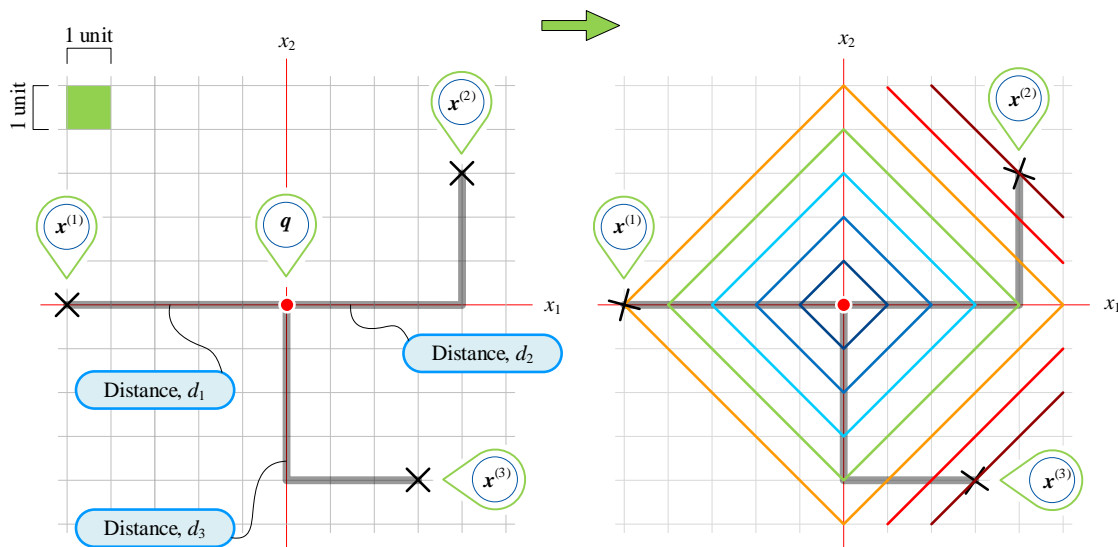
$$d(\mathbf{x}, \mathbf{q}) = |x_1 - q_1| + |x_2 - q_2| + \dots + |x_D - q_D| \quad (21)$$

特别地，当 $D = 2$ 时，城市街区距离为：

$$d(\mathbf{x}, \mathbf{q}) = |x_1 - q_1| + |x_2 - q_2| \quad (22)$$

旋转正方形

如图 10 所示，城市街区距离的等距线为旋转正方形。图中， $\mathbf{x}^{(1)}$ 、 $\mathbf{x}^{(2)}$ 和 $\mathbf{x}^{(3)}$ 和 \mathbf{q} 欧氏距离均为 5，但是城市街区距离分别为 5、7 和 7。

图 10.2 特征 ($D=2$) 城市街区距离

代码 Bk6_Ch05_05.ipynb 给出两种方法计算得到图 10 所示城市街区距离。

5.6 切比雪夫距离： L^∞ 范数

切比雪夫距离 (Chebyshev distance)，具体如下：

$$d(\mathbf{x}, \mathbf{q}) = \|\mathbf{x} - \mathbf{q}\|_\infty = \max_j \{|x_j - q_j|\} \quad (23)$$



切比雪夫距离就是我们在《矩阵力量》第 3 章中介绍的 L^∞ 范数。

将 (23) 展开得到下式：

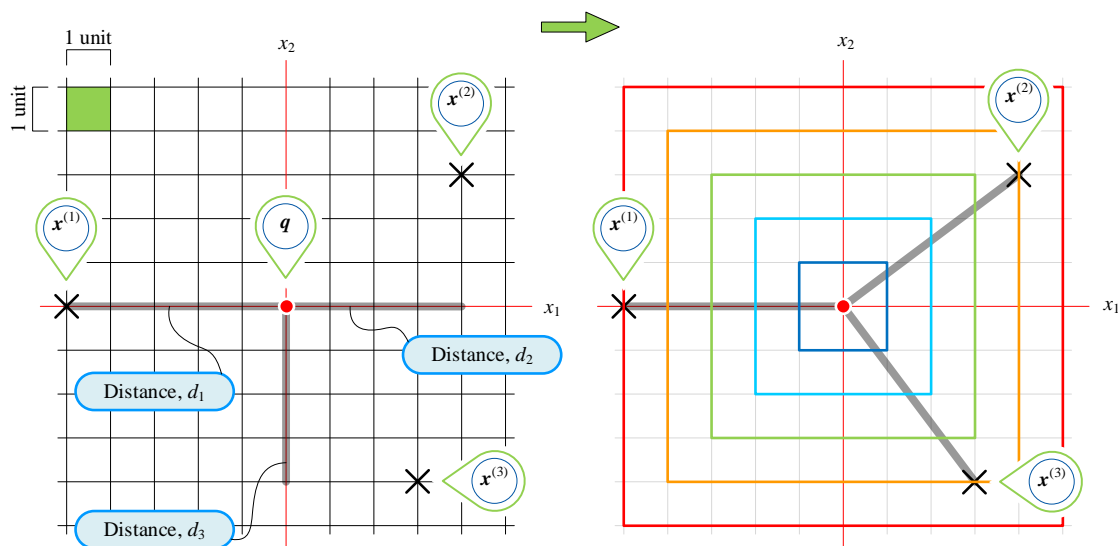
$$d(\mathbf{x}, \mathbf{q}) = \max \{|x_1 - q_1|, |x_2 - q_2|, \dots, |x_D - q_D|\} \quad (24)$$

特别地，当 $D=2$ 时，切比雪夫距离为：

$$d(\mathbf{x}, \mathbf{q}) = \max \{|x_1 - q_1|, |x_2 - q_2|\} \quad (25)$$

正方形

如图 11 所示，切比雪夫距离等距线为正方形。前文提到， $\mathbf{x}^{(1)}$ 、 $\mathbf{x}^{(2)}$ 和 $\mathbf{x}^{(3)}$ 和 \mathbf{q} 欧氏距离相同，但是切比雪夫距离分别为 5、4 和 4。

图 11.2 特征 ($D=2$) 切比雪夫距离

代码 Bk6_Ch05_06.ipynb 计算图 11 所示切比雪夫距离。

5.7 闵氏距离： L^p 范数

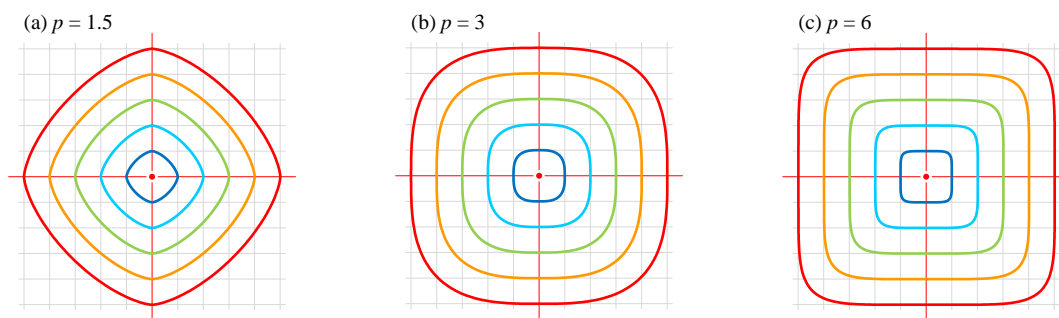
闵氏距离 (Minkowski distance) 类似 L^p 范数，对应定义如下：

$$d(\mathbf{x}, \mathbf{q}) = \|\mathbf{x} - \mathbf{q}\|_p = \left(\sum_{j=1}^D |x_j - q_j|^p \right)^{1/p} \quad (26)$$

⚠ 注意， $p \geq 1$ 时上式才叫向量范数。

计算闵氏距离的函数为 `scipy.spatial.distance.minkowski()`。

图 12 所示为 p 取不同值时，闵氏距离等距线图。特别地， $p=1$ 时，闵氏距离为城市街区距离； $p=2$ 时，闵氏距离为欧氏距离； $p \rightarrow \infty$ 时，闵氏距离为切比雪夫距离。

图 12. 闵氏距离 ($D=2$), p 取不同值

5.8 距离与亲近

本节介绍和距离相反的度量——**亲近度** (affinity)。两个样本数据距离越远，两者亲近度越低；而当它们距离越近，亲近度则越高。亲近度，也称**相似度** (similarity)。

余弦相似度

《矩阵力量》第 2 章讲过，**余弦相似度** (cosine similarity) 用向量夹角的余弦值度量样本数据的相似性。 \mathbf{x} 和 \mathbf{q} 两个向量的余弦相似度具体定义如下：

$$k(\mathbf{x}, \mathbf{q}) = \frac{\mathbf{x}^T \mathbf{q}}{\|\mathbf{x}\| \|\mathbf{q}\|} = \frac{\mathbf{x} \cdot \mathbf{q}}{\|\mathbf{x}\| \|\mathbf{q}\|} \quad (27)$$

如图 13 所示，如果两个向量方向相同，则夹角 θ 余弦值 $\cos(\theta)$ 为 1；如果，两个向量方向完全相反，夹角 θ 余弦值 $\cos(\theta)$ 为 -1。因此余弦相似度取值范围在 $[-1, +1]$ 之间。

⚠ 注意，余弦相似度和向量模无关，仅仅与两个向量夹角有关。

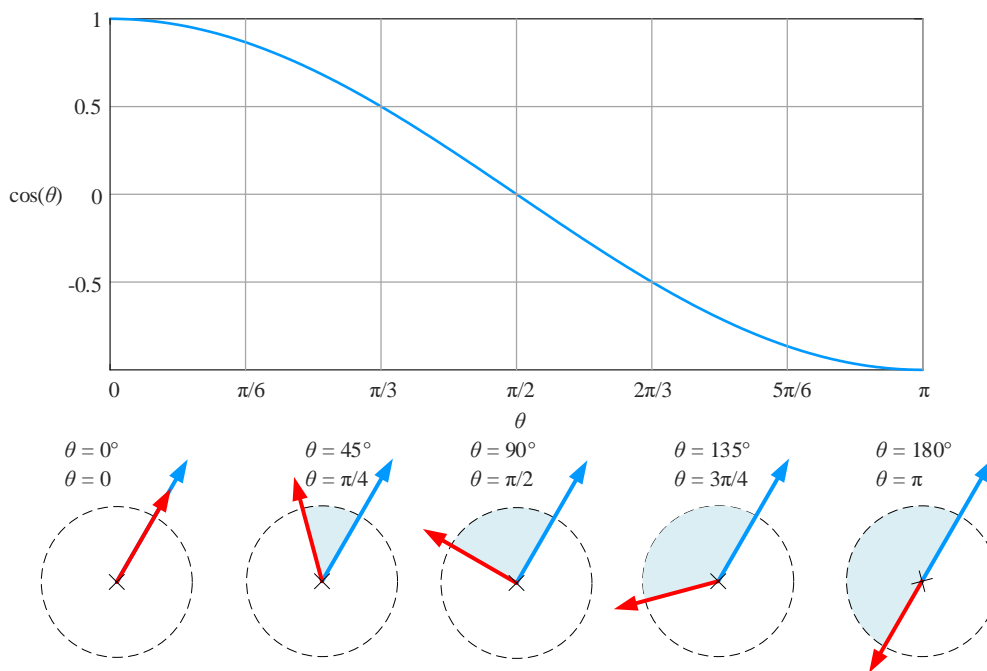


图 13. 余弦相似度

举个例子

给定如下两个向量具体值：

$$\mathbf{x} = [8 \ 2]^T, \quad \mathbf{q} = [7 \ 9]^T \quad (28)$$

将 (28) 代入 (27) 得到：

$$k(\mathbf{x}, \mathbf{q}) = \frac{\mathbf{x} \cdot \mathbf{q}}{\|\mathbf{x}\| \|\mathbf{q}\|} = \frac{8 \times 7 + 2 \times 9}{\sqrt{8^2 + 2^2} \times \sqrt{7^2 + 9^2}} = \frac{74}{\sqrt{68} \times \sqrt{130}} = 0.7871 \quad (29)$$



代码 Bk6_Ch05_07.ipynb 得到和 (29) 一致结果。

余弦距离

余弦距离 (cosine distance) 的定义如下：

$$d(\mathbf{x}, \mathbf{q}) = 1 - k(\mathbf{x}, \mathbf{q}) = 1 - \frac{\mathbf{x}^T \mathbf{q}}{\|\mathbf{x}\| \|\mathbf{q}\|} = 1 - \frac{\mathbf{x} \cdot \mathbf{q}}{\|\mathbf{x}\| \|\mathbf{q}\|} \quad (30)$$

余弦相似度的取值范围 $[-1, +1]$ 之间，因此余弦距离的取值范围为 $[0, 2]$ 。



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

Bk6_Ch05_08.ipynb 计算 (28) 中两个向量的余弦距离，结果为 0.2129。也可以采用 `scipy.spatial.distance.pdist(X, 'cosine')` 函数计算余弦距离。

相关系数相似度

相关系数相似度 (correlation similarity) 定义如下：

$$k(\mathbf{x}, \mathbf{q}) = \frac{(\mathbf{x} - \bar{\mathbf{x}})^T (\mathbf{q} - \bar{\mathbf{q}})}{\|\mathbf{x} - \bar{\mathbf{x}}\| \|\mathbf{q} - \bar{\mathbf{q}}\|} = \frac{(\mathbf{x} - \bar{\mathbf{x}}) \cdot (\mathbf{q} - \bar{\mathbf{q}})}{\|\mathbf{x} - \bar{\mathbf{x}}\| \|\mathbf{q} - \bar{\mathbf{q}}\|} \quad (31)$$

其中， $\bar{\mathbf{x}}$ 为列向量 \mathbf{x} 元素均值； $\bar{\mathbf{q}}$ 为列向量 \mathbf{q} 元素均值。

观察 (31)，发现相关系数相似度类似余弦相似度；稍有不同的是，相关系数相似度需要“中心化”向量。

还是以 (28) 为例，计算 \mathbf{x} 和 \mathbf{q} 两个向量的相关系数相似度。将 (28) 代入 (31) 可以得到：

$$\begin{aligned} k(\mathbf{x}, \mathbf{q}) &= \frac{\left(\begin{bmatrix} 8 & 2 \end{bmatrix}^T - \frac{8+2}{2} \right) \cdot \left(\begin{bmatrix} 7 & 9 \end{bmatrix}^T - \frac{7+9}{2} \right)}{\left\| \begin{bmatrix} 8 & 2 \end{bmatrix}^T - \frac{8+2}{2} \right\| \left\| \begin{bmatrix} 7 & 9 \end{bmatrix}^T - \frac{7+9}{2} \right\|} \\ &= \frac{\begin{bmatrix} 3 & -3 \end{bmatrix}^T \cdot \begin{bmatrix} -1 & 1 \end{bmatrix}^T}{\left\| \begin{bmatrix} 3 & -3 \end{bmatrix}^T \right\| \left\| \begin{bmatrix} -1 & 1 \end{bmatrix}^T \right\|} = \frac{-6}{6} = -1 \end{aligned} \quad (32)$$



代码 Bk6_Ch05_09.ipynb 计算得到两个向量的相关系数距离为 2。也可以采用 `scipy.spatial.distance.pdist(X, 'correlation')` 函数计算相关系数距离。

核函数亲近度

不考虑常数项，**线性核** (linear kernel) 亲近度定义如下：

$$\kappa(\mathbf{x}, \mathbf{q}) = \mathbf{x}^T \mathbf{q} = \mathbf{x} \cdot \mathbf{q} \quad (33)$$

对比 (27) 和 (33)，(27) 分母上 $\|\mathbf{x}\|$ 和 $\|\mathbf{q}\|$ 分别对 \mathbf{x} 和 \mathbf{q} 归一化。

`sklearn.metrics.pairwise.linear_kernel` 为 scikit-learn 工具箱中计算线性核亲近度函数。

将 (28) 代入 (33)，得到线性核亲近度为：

$$\kappa(\mathbf{x}, \mathbf{q}) = 8 \times 7 + 2 \times 9 = 74 \quad (34)$$

多项式核 (polynomial kernel) 亲近度定义如下：

$$\kappa(\mathbf{x}, \mathbf{q}) = (\gamma \mathbf{x}^T \mathbf{q} + r)^d = (\gamma \mathbf{x} \cdot \mathbf{q} + r)^d \quad (35)$$

其中， d 为多项式核次数， γ 为系数， r 为常数。

多项式核亲近度函数为 `sklearn.metrics.pairwise.polynomial_kernel`。

Sigmoid 核 (sigmoid kernel) 亲近度定义如下：

$$\kappa(\mathbf{x}, \mathbf{q}) = \tanh(\gamma \mathbf{x}^T \mathbf{q} + r) = \tanh(\gamma \mathbf{x} \cdot \mathbf{q} + r) \quad (36)$$

Sigmoid 核亲近度函数为 `sklearn.metrics.pairwise.sigmoid_kernel`。

最常见的莫过于，**高斯核** (Gaussian kernel) 亲近度，即**径向基核函数** (radial basis function kernel, RBF kernel)：

$$\kappa(\mathbf{x}, \mathbf{q}) = \exp(-\gamma \|\mathbf{x} - \mathbf{q}\|^2) \quad (37)$$

(37) 中 $\|\mathbf{x} - \mathbf{q}\|^2$ 为欧氏距离的平方，(37) 也可以写作：

$$\kappa(\mathbf{x}, \mathbf{q}) = \exp(-\gamma d^2) \quad (38)$$

其中， d 为欧氏距离 $\|\mathbf{x} - \mathbf{q}\|$ 。高斯核亲近度取值范围为 $(0, 1]$ ；距离值越小，亲近度越高。高斯核亲近度函数为 `sklearn.metrics.pairwise.rbf_kernel`。

图 14 所示为， γ 取不同值时，高斯核亲近度随着欧氏距离 d 变化。聚类算法经常采用高斯核亲近度。

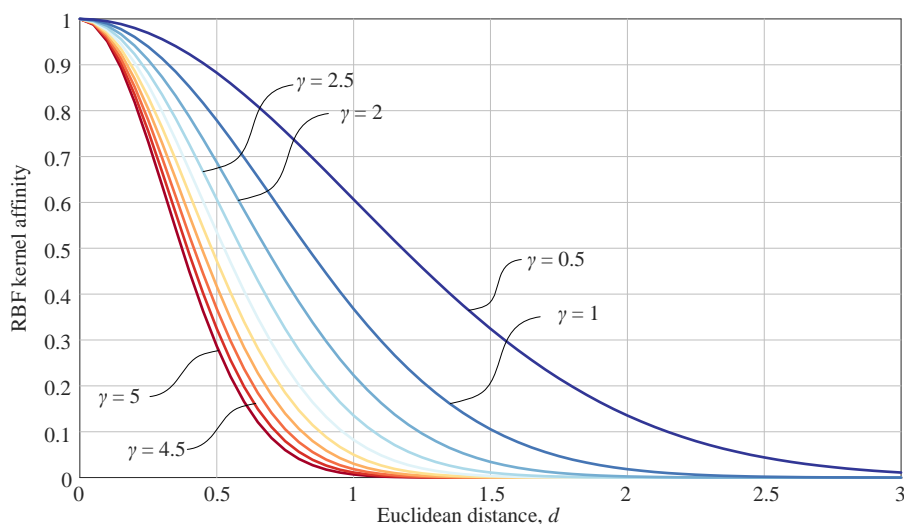


图 14. 高斯核亲近度随欧氏距离变化

从“距离 → 亲近度”转换角度来看，多元高斯分布分子中高斯函数完成的的就是马氏距离 d 到概率密度 (亲近度) 的转化：

$$f_{\mathbf{x}}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} = \frac{\exp\left(-\frac{1}{2}d^2\right)}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \quad (39)$$

拉普拉斯核 (Laplacian kernel) 亲密度，定义如下：

$$\kappa(\mathbf{x}, \mathbf{q}) = \exp(-\gamma \|\mathbf{x} - \mathbf{q}\|_1) \quad (40)$$

其中， $\|\mathbf{x} - \mathbf{q}\|_1$ 为城市街区距离。

图 15 所示为， γ 取不同值时，拉普拉斯核亲密度随着城市街区距离 d 变化。拉普拉斯核亲密度对应函数为 `sklearn.metrics.pairwise.laplacian_kernel`。

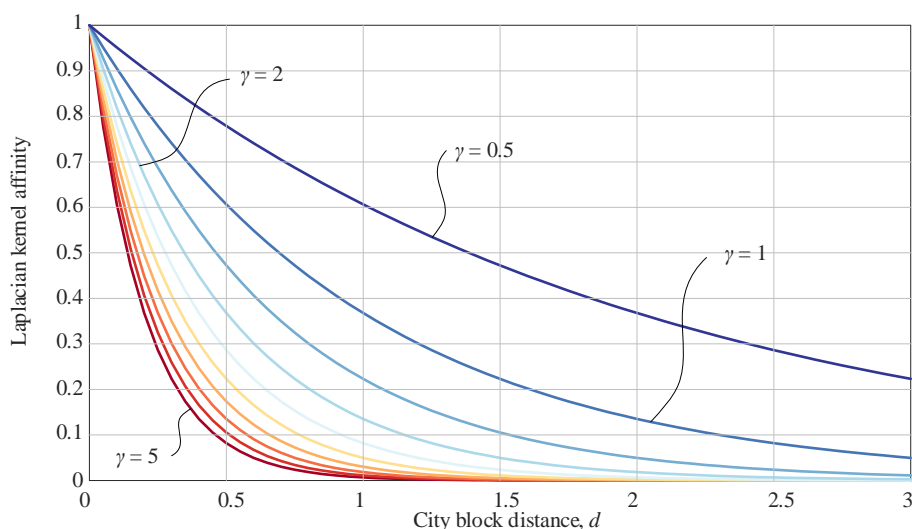


图 15. 拉普拉斯核亲密度随距离变化

5.9 成对距离、成对亲密度

《矩阵力量》反复强调，样本数据矩阵 \mathbf{X} 每一列代表一个特征，而每一行代表一个样本数据点，比如：

$$\mathbf{X}_{n \times D} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix} \quad (41)$$

\mathbf{X} 样本点之间距离构成的**成对距离矩阵** (pairwise distance matrix) 形式如下：

$$\mathbf{D}_{n \times n} = \begin{bmatrix} 0 & d_{1,2} & d_{1,3} & \cdots & d_{1,n} \\ d_{2,1} & 0 & d_{2,3} & \cdots & d_{2,n} \\ d_{3,1} & d_{3,2} & 0 & \cdots & d_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n,1} & d_{n,2} & d_{n,3} & \cdots & 0 \end{bmatrix} \quad (42)$$

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

每个样本数据点和自身的距离为 0，因此 (42) 主对角线为 0。很显然矩阵 D 为对称矩阵，即 d_{ij} 和 d_{ji} 相等。

图 16 给定 12 个样本数据点坐标点。

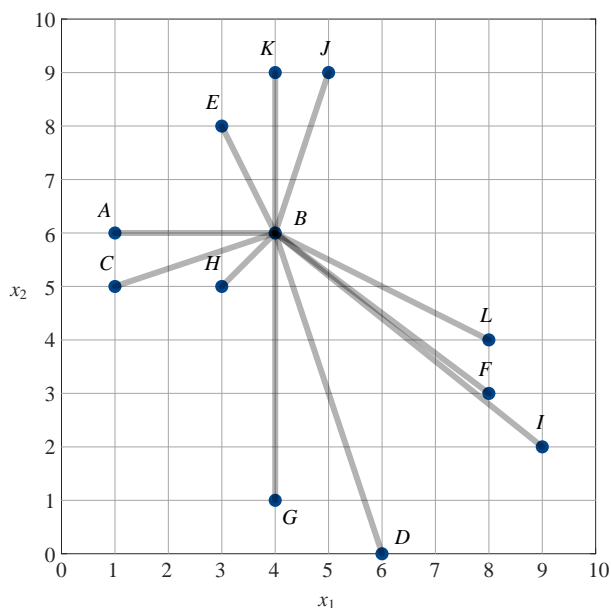


图 16. 样本数据散点图和成对距离

利用 `sklearn.metrics.pairwise.euclidean_distances`，我们可以计算图 16 数据点的成对欧氏距离矩阵。图 17 所示为欧氏距离矩阵数据构造的热图。

实际上，我们关心的成对距离个数为：

$$C_n^2 = \frac{n(n-1)}{2} \quad (43)$$

也就是说，(42) 中不含对角线的下三角矩阵包含的信息足够使用。

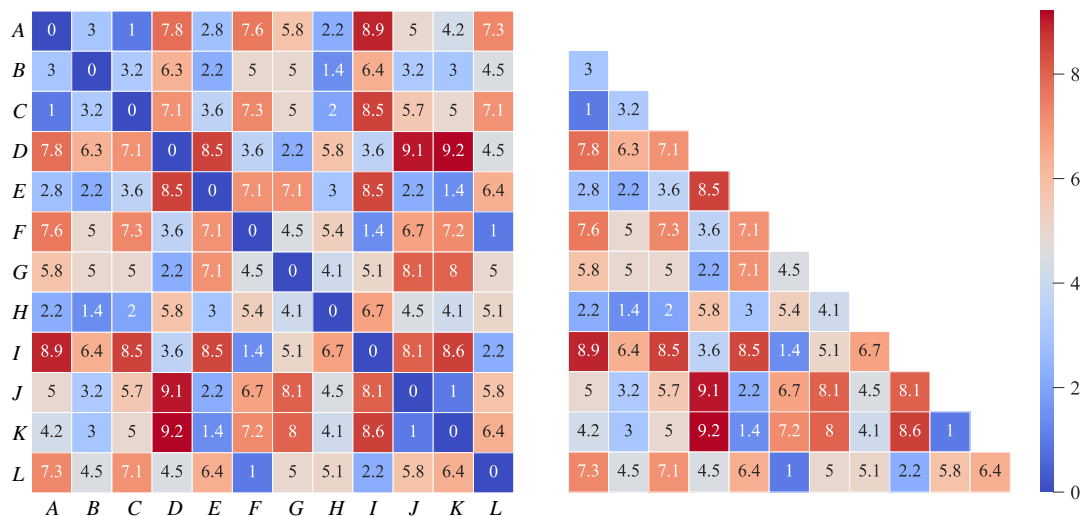


图 17. 样本数据成对距离矩阵热图

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

表 1 总结计算成对距离、亲密度矩阵常用函数。

表 1. 计算成对距离/亲密度矩阵常见函数

函数	描述
<code>metrics.pairwise.cosine_similarity()</code>	计算余弦相似度成对矩阵
<code>metrics.pairwise.cosine_distances()</code>	计算成对相似性距离矩阵
<code>metrics.pairwise.euclidean_distances()</code>	计算成对欧氏距离矩阵
<code>metrics.pairwise.laplacian_kernel()</code>	计算拉普拉斯核成对亲密度矩阵
<code>metrics.pairwise.linear_kernel()</code>	计算线性核成对亲密度矩阵
<code>metrics.pairwise.manhattan_distances()</code>	计算成对城市街区距离矩阵
<code>metrics.pairwise.polynomial_kernel()</code>	计算多项式核成对亲密度矩阵
<code>metrics.pairwise.rbf_kernel()</code>	计算 RBF 核成对亲密度矩阵
<code>metrics.pairwise.sigmoid_kernel()</code>	计算 sigmoid 核成对亲密度矩阵
<code>metrics.pairwise.paired_euclidean_distances(X,Q)</code>	计算 X 和 Q 样本数据矩阵成对欧氏距离矩阵
<code>metrics.pairwise.paired_manhattan_distances(X,Q)</code>	计算 X 和 Q 样本数据矩阵成对城市街区距离矩阵
<code>metrics.pairwise.paired_cosine_distances(X,Q)</code>	计算 X 和 Q 样本数据矩阵成对余弦距离矩阵



代码 Bk6_Ch05_10.ipynb 可以绘制图 16、图 17。

5.10 协方差矩阵，为什么无处不在？

想要可视化一个 n 行 D 列的数据矩阵 X ，成对散点图是个不错的选择。图 18 所示为用 `seaborn.pairplot()` 绘制的成对散点图。这幅图有 D 行、 D 列个子图，其实也可以看成是个方阵。

对角线上的子图展示的是概率密度曲线，在这些图中我们可以看到不同特征有不同分布特点；非对角线子图展示的是成对散点图，这些子图中我们似乎看到某些散点子图有更强的正相关性。

那么问题来了，如何量化上述观察？

协方差矩阵 (covariance matrix) 就派上了用场！

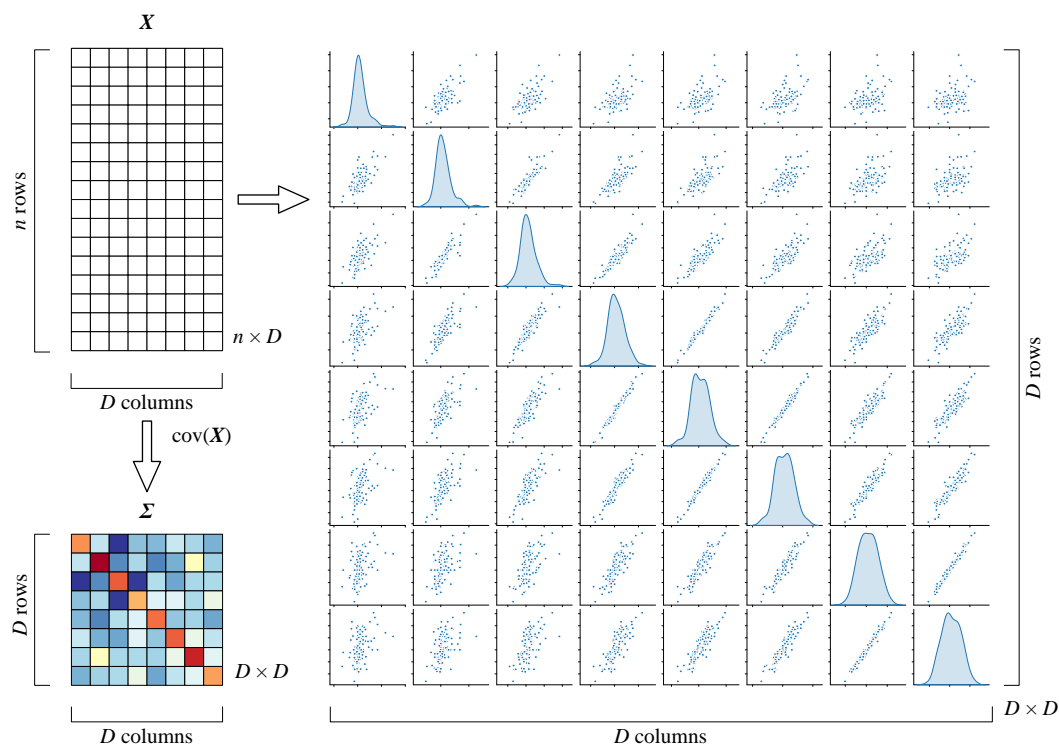


图 18. 成对散点图

观察图 18 这个协方差矩阵 Σ ，我们可以发现 Σ 好像是个“浓缩”的成对散点图，它们的形状都是 $D \times D$ 。也就是说，成对散点图的每个子图浓缩成了 Σ 中的一个值。

协方差矩阵 Σ 主对角线为**方差** (variance)，对应成对散点图中的主对角线子图，量化某个特定特征上样本数据分布离散情况。 $D \times D$ 协方差矩阵有 D 个方差。本章前文提到方差也相当于某种距离。

协方差矩阵 Σ 非主对角线为**协方差** (covariance)，对应成对散点图中的非主对角线子图，量化成对特征的关系。 $D \times D$ 协方差矩阵有 $D^2 - D = D(D - 1)$ 个协方差。协方差度量特征之间相关性强度，某种程度上也可以视作“距离”。

更何况，协方差矩阵直接用在马氏距离计算中。

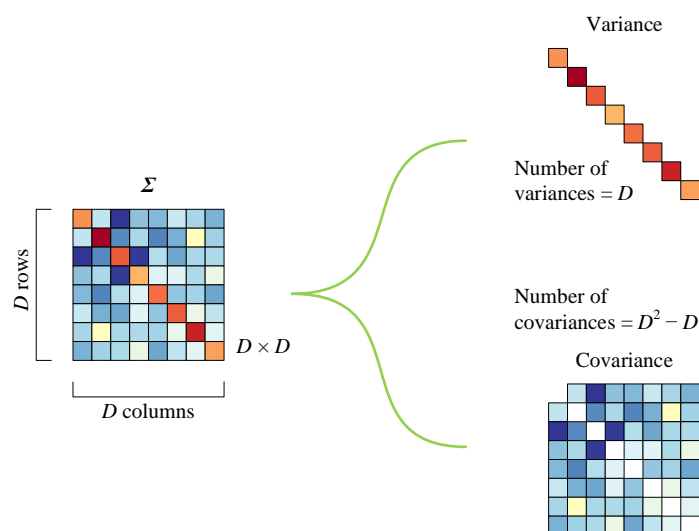


图 19. 协方差矩阵由方差和协方差组成

由于计算协方差矩阵时，每个特征上的数据都已经去均值，因此 Σ 不含有 X 的质心 $E(X)$ 具体信息。

对于鸢尾花书的读者，“协方差矩阵”这个词可能已经给大家的耳朵磨出茧子。本书经常提到协方差矩阵，是因为机器学习很多算法都离不开协方差矩阵。

首先，协方差矩阵直接用在**多元高斯分布** (multivariate Gaussian distribution) PDF 中。**马氏距离** (Mahal distance, Mahalanobis distance) 也离不开协方差矩阵；除了作为距离度量，马氏距离常常用来判断**离群值** (outlier)。

条件高斯分布 (conditional Gaussian distribution) 也离不开协方差矩阵的分块运算。而条件高斯分布常用在多输入多输出的线性回归中；此外，我们将会在**高斯过程** (Gaussian Process, GP) 中用到高斯条件概率。特别对于高斯过程算法，我们要用不同的核函数构造先验分布的协方差矩阵。

在主成分分析 (Principal Component Analysis, PCA) 中，一般都是以特征值分解协方差矩阵为起点。PCA 的主要思想是找到数据中的主成分，这些主成分是原始特征的线性组合。协方差矩阵用于计算数据的特征向量和特征值，特征向量构成了新的坐标系，而特征值表示了每个主成分的重要性。

在**高斯混合模型** (Gaussian Mixture Model, GMM) 中，每个混合成分都由一个高斯分布表示，而每个高斯分布都有一个协方差矩阵。协方差矩阵决定了每个混合成分在特征空间中的形状和方向。不同的协方差矩阵可以捕捉到不同方向上的数据变化。

高斯朴素贝叶斯 (Gaussian Naive Bayes) 算法中，每个类别的特征都被假设为服从高斯分布。协方差矩阵描述每个类别中不同特征之间关系。该方法假设每个类别下的协方差矩阵为对角阵，即特征之间的关系是条件独立的，因此被称为“朴素”。

高斯判别分析 (Gaussian Discriminant Analysis, GDA) 是一种监督学习算法，通常用于分类问题。GDA 使用协方差矩阵来建模每个类别的特征分布。与高斯朴素贝叶斯不同，GDA 中协方差矩阵未必假定是对角矩阵，因此能够捕捉到不同特征之间的相关性。

当然协方差矩阵也不是万能的！

协方差矩阵通常假设数据服从多元高斯分布。如果数据的分布不符合这个假设，协方差矩阵可能不是一个有效的描述统计关系的工具。如果数据分布呈现偏斜或非正态分布，协方差矩阵的解释力可能受到影响。在这种情况下，可能需要考虑对数据进行转换或使用其他方法。

协方差矩阵受到特征的取值尺度、单位等影响。为了解决这个问题，我们可以采用相关性系数矩阵，即原始数据 z 分数的协方差矩阵。

协方差受异常值的影响较大，如果数据中存在离群值，协方差矩阵可能不够稳健。

协方差矩阵主要用于捕捉线性关系，对于非线性关系，协方差矩阵可能无法提供很好的信息。在这种情况下，非线性方法或核方法可能更适用。

随着特征数量的增加，协方差矩阵的计算和存储成本会显著增加。当特征维度很高时，计算协方差矩阵可能变得非常耗时，并且需要更多的内存。

本章后文一边回顾鸢尾花书前五本书介绍的有关协方差的重要知识点，然后再扩展讲解一些新内容。

怎么计算数据的协方差矩阵？

相信大家已经很熟悉计算协方差矩阵 Σ 的具体步骤，下面简单回顾。

如图 20 所示，对于原始数据矩阵 X ，首先对其中心化得到 X_c 。从几何角度来看，中心化相当于平移，将质心从 $E(X)$ 平移到原点。

然后计算 X_c 的格拉姆矩阵 $X_c^T X_c$ ，并用 $1/(n-1)$ 缩放。

$$\Sigma = \frac{X_c^T X_c}{n-1} \quad (44)$$

如果假设 X 已经标准化，协方差矩阵可以简单写成 $\Sigma = \frac{X^T X}{n-1}$ ；也就是可以这样理解，协方差矩阵 Σ 是一种特殊的。

很多时候，特别是对协方差矩阵 Σ 特征值分解，我们甚至可以不考虑缩放系数 $1/(n-1)$ 。图 20 中，如果将 Demean 改成 Standardize (标准化)，我们便得到的是相关性系数矩阵 P 。或者说， X 的 z 分数矩阵的协方差矩阵就是 X 的**相关性系数矩阵** (correlation matrix)。相关性系数矩阵的主对角线元素都为 1，非主对角线元素为相关性系数。

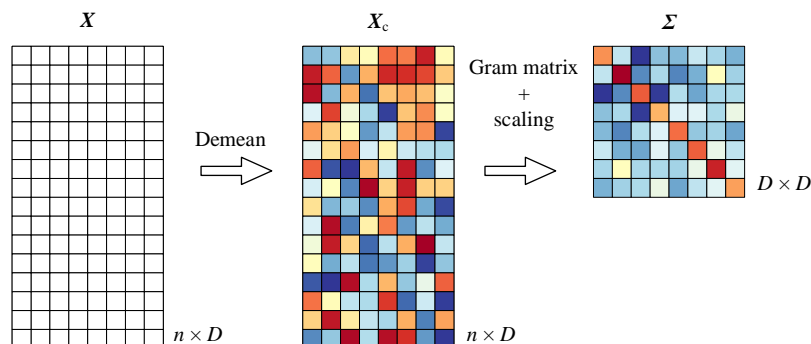


图 20. 计算协方差矩阵

相对于形状为 $n \times D$ 的数据矩阵 X ，一般情况 $n \gg D$ ，即 n 远大于 D ，一个 $D \times D$ 的协方差矩阵 Σ 则小巧轻便的多。 Σ 不但包含 X 每一列数据的方差，还包含 X 任意两列数据的协方差。

矮胖矩阵的协方差矩阵

前文的数据矩阵形状都是细高，即矩阵的行数 n 大于列数 D 。但是，实践中，我们也会经常碰到矮胖型的数据矩阵，即 $n < D$ 。比如，2000 (D) 只股票在 252 (n) 个交易日的数据。

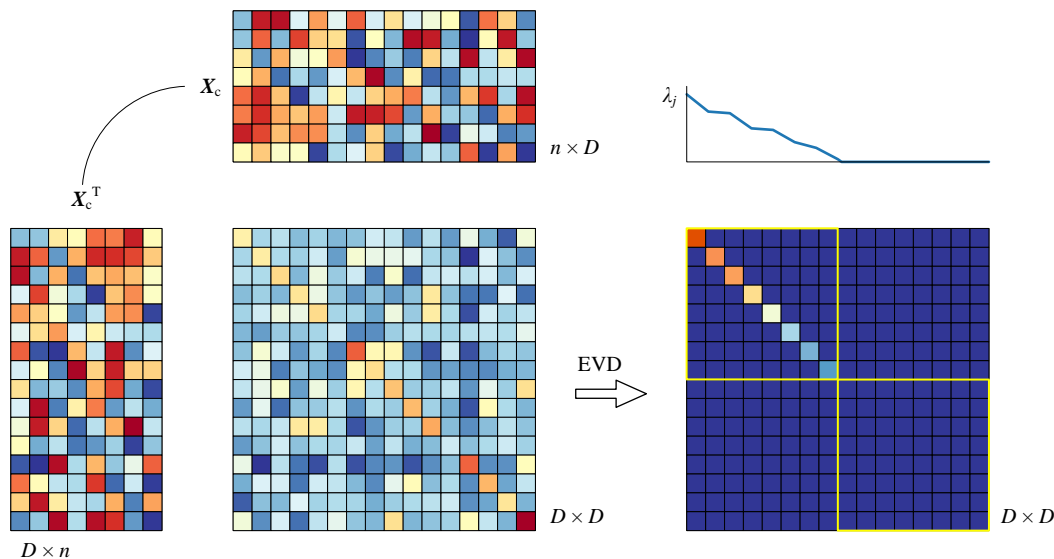


图 21. 协方差矩阵存在大量 0 特征值

如图 21 所示，对于矮胖数据矩阵的协方差矩阵，它的秩远小于 D ；可以肯定地说，这种协方差矩阵一定是半正定，即不能进行 Cholesky 分解。此外，对图 21 中协方差矩阵特征值分解时，我们会看到大量特征值为 0，这会造成运算不稳定。这种情况下，我们可以将原始数据转置后再计算“细高”矩阵的协方差矩阵，然后再进行矩阵分解（特征值分解、Cholesky 分解等）。

矩阵乘法两个视角

下面用矩阵乘法两个视角来观察 (44)。

根据矩阵乘法第一视角，将 X_c 写成 $[x_1 \ x_2 \ \dots \ x_D]$ ，(44) 可以展开写成。

$$\text{var}(X) = \Sigma = \frac{1}{n-1} \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \dots & x_1^T x_D \\ x_2^T x_1 & x_2^T x_2 & \dots & x_2^T x_D \\ \vdots & \vdots & \ddots & \vdots \\ x_D^T x_1 & x_D^T x_2 & \dots & x_D^T x_D \end{bmatrix} = \frac{1}{n-1} \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \dots & \langle x_1, x_D \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \dots & \langle x_2, x_D \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_D, x_1 \rangle & \langle x_D, x_2 \rangle & \dots & \langle x_D, x_D \rangle \end{bmatrix} \quad (45)$$

注意，上式中 x_j ($j = 1, 2, \dots, D$) 已经中心化，即去均值。

如图 22 所示，协方差矩阵的主对角线元素为 $x_j^T x_j$ ，相当于向量内积 $\langle x_j, x_j \rangle$ ，也相当于向量 x_j 的 L2 范数平方 $\|x_j\|_2^2$ 。

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

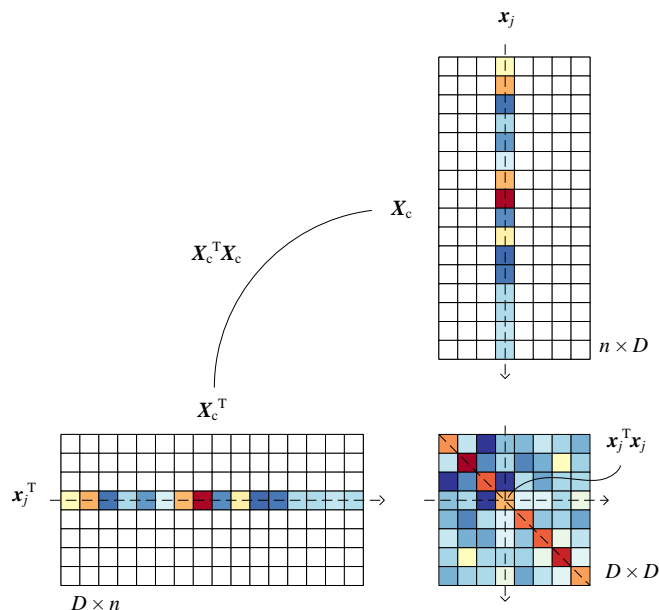


图 22. 协方差矩阵主对角线元素

如图 24 所示，协方差矩阵的非主对角线元素为 $\mathbf{x}_j^T \mathbf{x}_k$ ($j \neq k$)，相当于向量内积 $\langle \mathbf{x}_j, \mathbf{x}_k \rangle$ 。显然， $\mathbf{x}_j^T \mathbf{x}_k = \mathbf{x}_k^T \mathbf{x}_j$ ，即 $\langle \mathbf{x}_j, \mathbf{x}_k \rangle = \langle \mathbf{x}_k, \mathbf{x}_j \rangle$ ；也就是说，协方差矩阵为**对称矩阵** (symmetric matrix)。

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 23 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，我们可以发现 $C_D^2 = \frac{D(D-1)}{2}$ ，表示在 D 个特征中任意取 2 个特征的组合数。

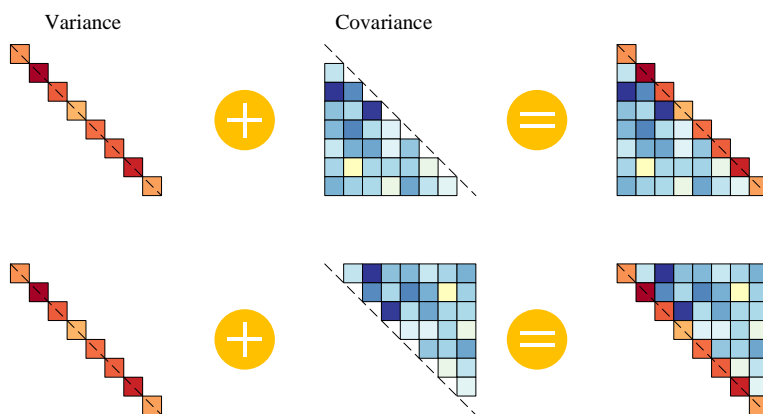


图 23. 剔除协方差矩阵中冗余元素

而根据方差非负这个形式，很容易证明对于非零向量 \mathbf{a} ， $\mathbf{a}^T \Sigma \mathbf{a} \geq 0$ 成立；这也意味着，协方差矩阵为**半正定** (Positive semidefinite, PSD)。

对协方差矩阵 Σ 进行谱分解 $\Sigma = \mathbf{V} \Lambda \mathbf{V}^T$ ，如果得到的所有特征值 λ_j 均为正，则协方差矩阵正定；这也说明，数据矩阵满秩，即线性独立。如果协方差矩阵的特征值出现 0，就意味着 Σ 非满秩，也说明数据矩阵非满秩，存在线性相关。这一点值得我们注意，因为 Σ 非满秩，则意味着 Σ 不存在逆，行列式 $|\Sigma|$ 为 0。多元高斯分布 PDF 函数中， Σ 必须为正定。

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

从上面这些分析，也可以联想到为什么我们常常把线性代数中的矩阵形状、秩、矩阵逆、行列式、正定性、特征值等概念联系起来。

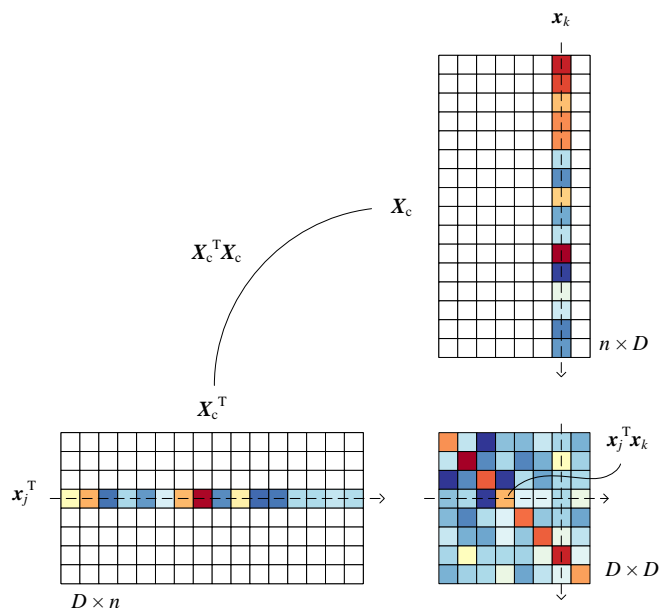


图 24. 协方差矩阵非主对角线元素

根据矩阵乘法第二视角，将 \mathbf{X}_c 写成 $\begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(n)} \end{bmatrix}$ ，(44) 可以展开写成 n 个秩一矩阵之和。

$$\Sigma = \frac{1}{n-1} \left[(\mathbf{x}^{(1)})^T \mathbf{x}^{(1)} + (\mathbf{x}^{(2)})^T \mathbf{x}^{(2)} + \dots + (\mathbf{x}^{(n)})^T \mathbf{x}^{(n)} \right] = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}^{(i)})^T \mathbf{x}^{(i)} \quad (46)$$

其中，每个 $(\mathbf{x}^{(i)})^T \mathbf{x}^{(i)}$ 均为**秩一矩阵** (rank-one matrix)，形状为 $D \times D$ 。

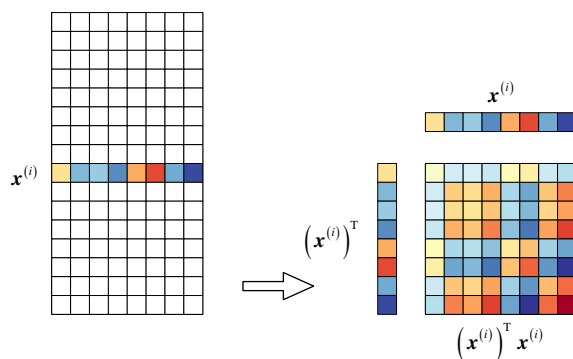


图 25. 协方差矩阵可以看成 n 个秩一矩阵之和

如图 26 所示，(46) 相当于对于 n 个 $(\mathbf{x}^{(i)})^T \mathbf{x}^{(i)}$ 取均值；而且，每个样本点都有相同的权重 $\frac{1}{n-1}$ 。

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

虽然 $(\mathbf{x}^{(i)})^T \mathbf{x}^{(i)}$ 的秩为 1，但是协方差矩阵 Σ 的秩最大为 D ， $\text{rank}(\Sigma) \leq D$ 。

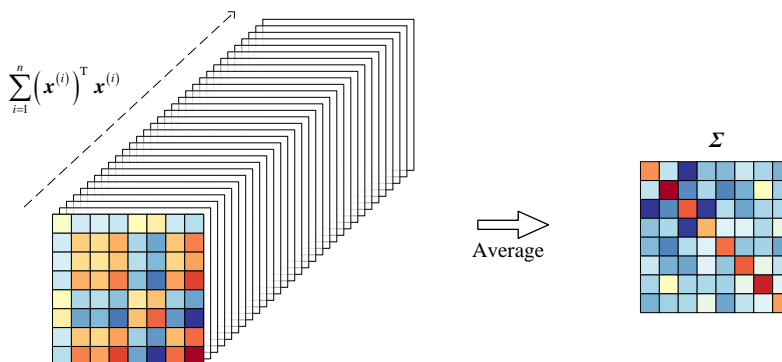


图 26. 协方差矩阵可以看成 n 个秩一矩阵取平均

几何视角：椭圆和椭球

如图 27 所示，任意 2×2 协方差矩阵可以看做是一个椭圆。椭圆的中心位于质心。

如图 28 所示，这个椭圆的形状和旋转角度则由相关系数和方差比值共同决定。请大家注意，图 28 中旋转椭圆都对应马氏距离为 1。《统计至简》还介绍了，条件高斯概率和这些图之间的关系，请大家自行回顾。

要求得椭圆的长轴、短轴各自所在方向，我们需要特征值分解协方差矩阵。

对协方差进行特征值分解时，获得的特征值大小和半长轴、半短轴长度直接相关。这实际上也是利用特征值分解完成 PCA 的几何解释。《矩阵力量》和《统计至简》都从不同角度介绍过相关内容，这里不再重复。

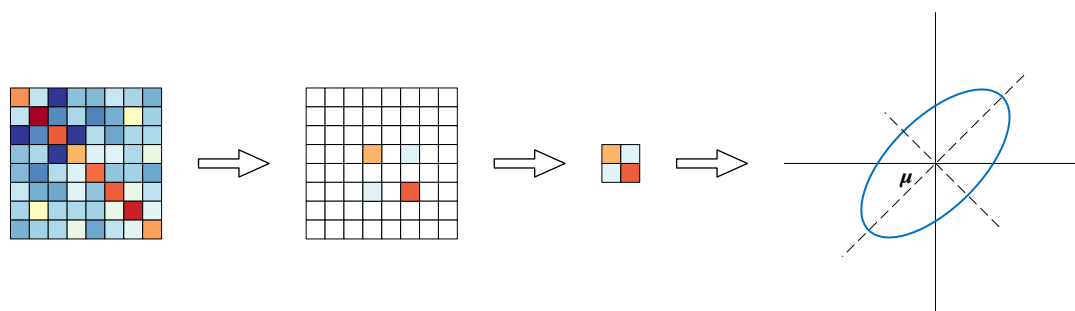
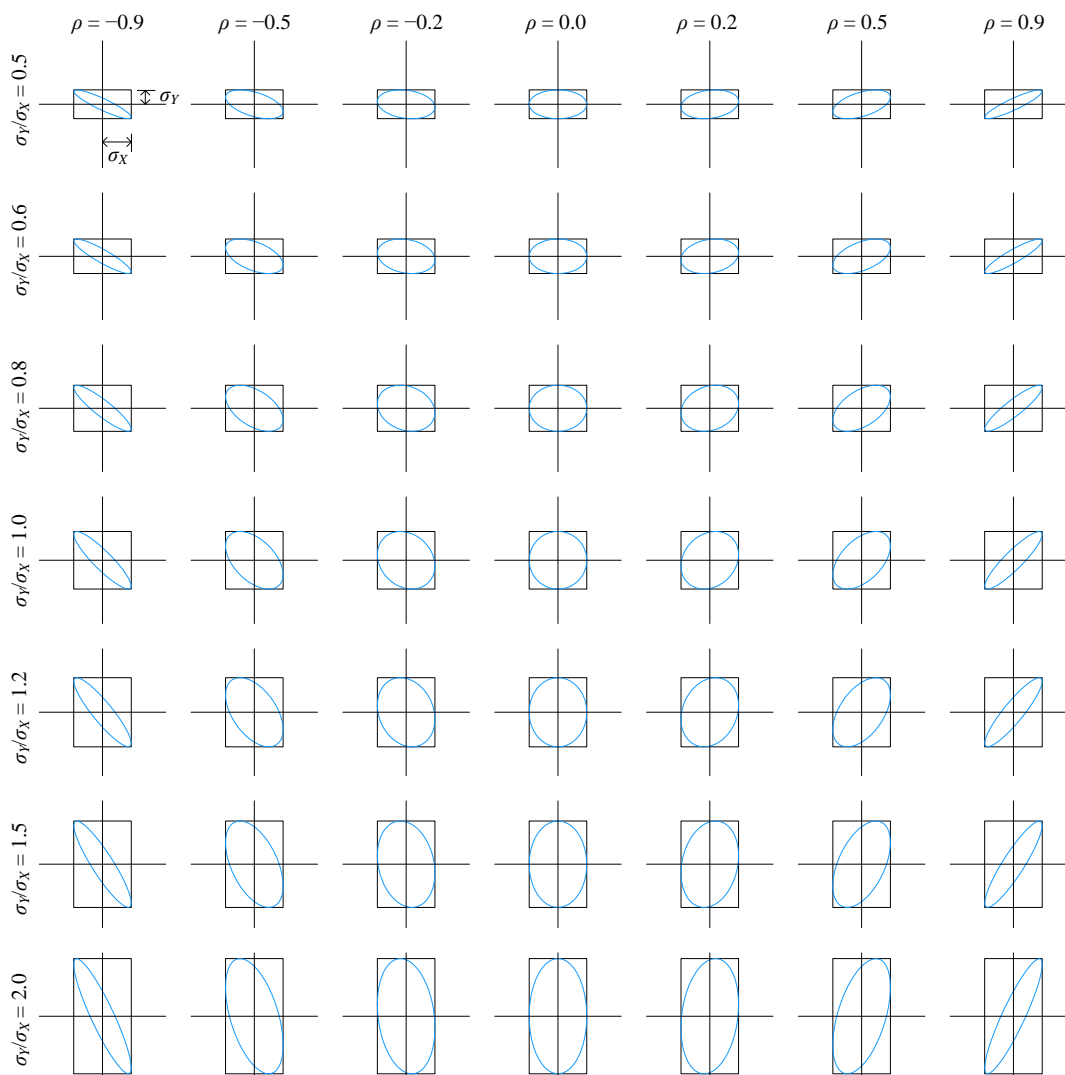


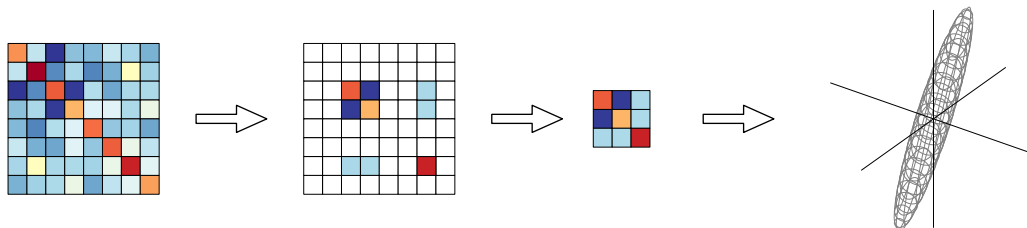
图 27. 任意 2×2 协方差矩阵可以看做是一个椭圆

图 28. 2×2 协方差椭圆随相关性系数 ρ 、标准差比值 σ_y/σ_x 变化

如图 29 所示，任意 3×3 协方差矩阵可以看做是一个椭球；这个椭球也对应马氏距离为 1。如图 30 所示，将这个椭球投影到三个平面上，我们便得到了三个椭圆，它们也是对应马氏距离为 1。我们可以用这三个椭圆代表三个不同的 2×2 协方差矩阵。

仔细观察图 30 中这个旋转椭球，我们还看到了三个向量。这三个向量分别代表椭球三个主轴方向。类似地，对这个 3×3 协方差矩阵进行特征值分解便可以获得这三个方向。

在《矩阵力量》中，我们知道这三个方向也是一个正交基。如图 31 所示，顺着这三个方向，我们可以把椭球摆正！

图 29. 任意 3×3 协方差矩阵可以看做是一个椭球

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

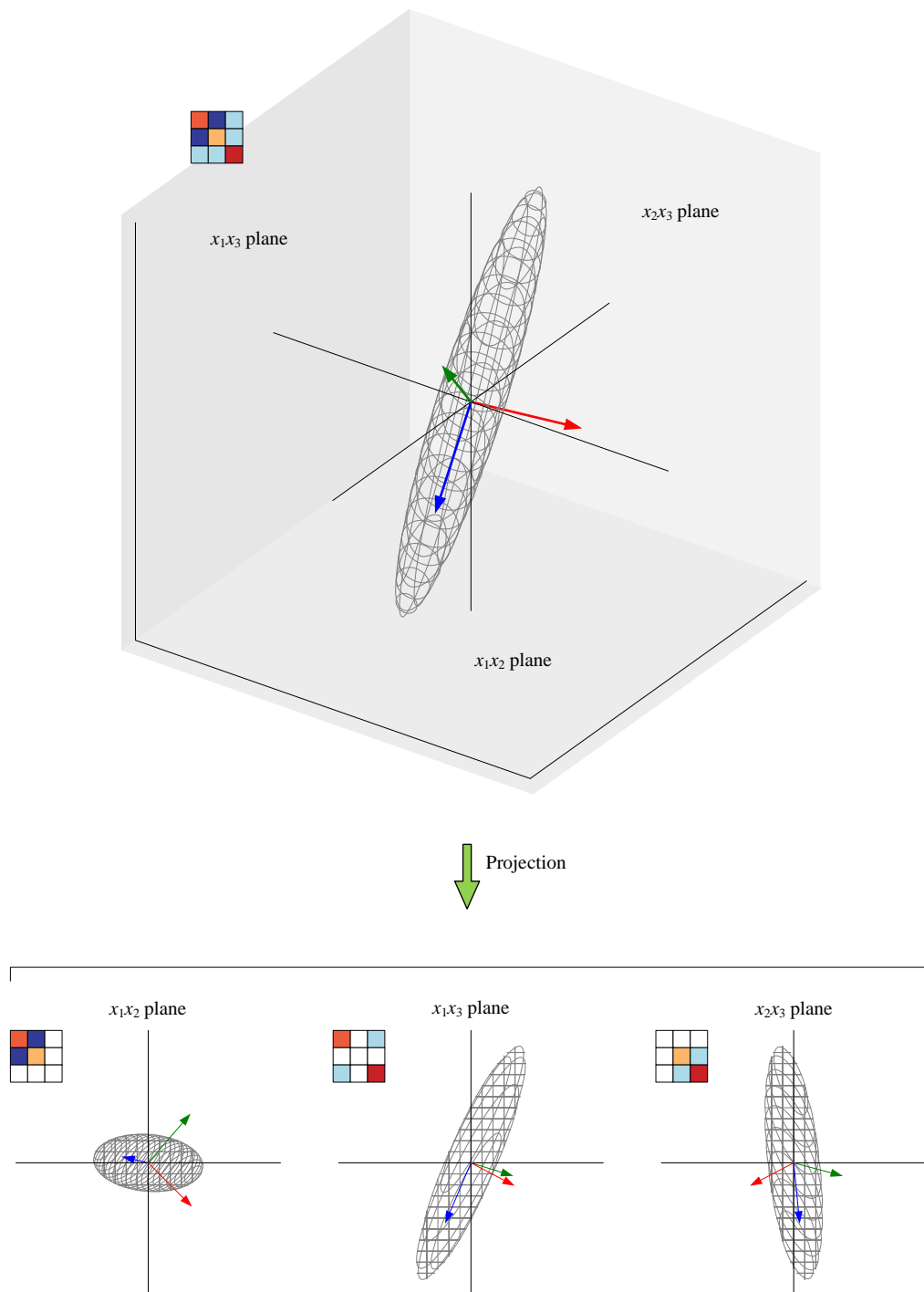


图 30. 椭球在三个平面的投影

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

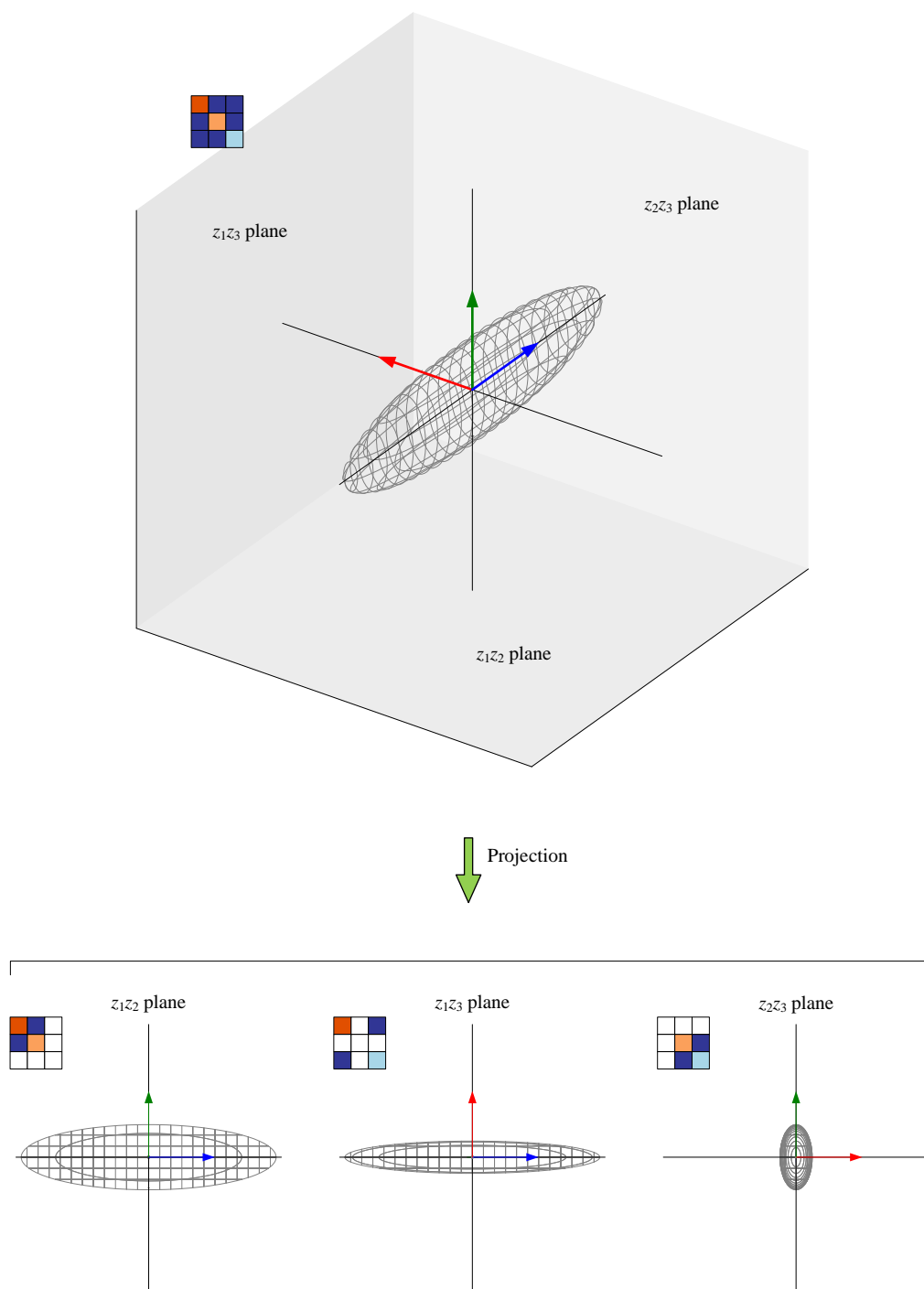


图 31. 把椭球摆正

谱分解：特征值分解特例

图 32 所示为协方差矩阵的谱分解。注意， V 为正交矩阵，即满足 $V^T V = V V^T = I$ 。

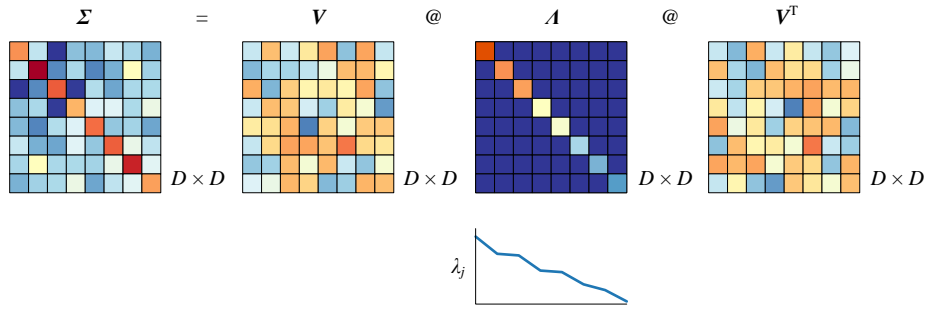


图 32. 协方差矩阵的谱分解

用类似方法，将谱分解结果 $\Sigma = V\Lambda V^T$ 展开为 D 个秩一矩阵相加。

$$\Sigma = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \lambda_D \mathbf{v}_D \mathbf{v}_D^T = \sum_{j=1}^D \lambda_j \mathbf{v}_j \mathbf{v}_j^T \quad (47)$$

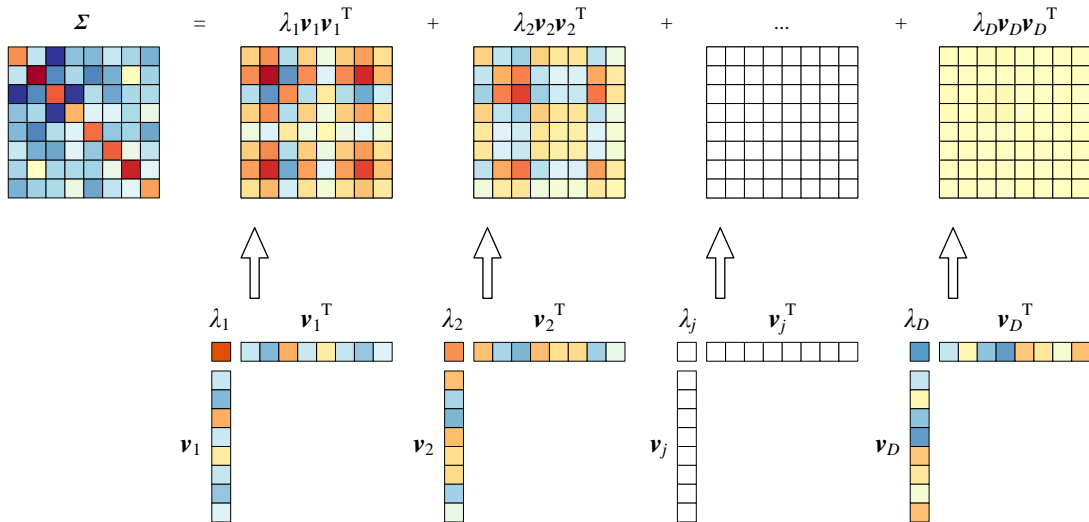
其中， $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D$ 。 $\lambda_j \mathbf{v}_j \mathbf{v}_j^T$ 也都是秩一矩阵。

此外， $\text{trace}(\Lambda) = \text{trace}(\Sigma)$ ，即 $\sum_{j=1}^D \lambda_j = \sum_{j=1}^D \sigma_j^2$ 。

由于 V 为正交矩阵，显然，当 $j \neq k$ 时， \mathbf{v}_j 和 \mathbf{v}_k 相互垂直，即 $\mathbf{v}_j^T \mathbf{v}_k = \mathbf{v}_j^T \mathbf{v}_k = 0$ ，也就是说 $\langle \mathbf{v}_j, \mathbf{x}_k \rangle = \langle \mathbf{v}_k, \mathbf{v}_j \rangle = 0$ 。而投影矩阵 $\mathbf{v}_j \mathbf{v}_j^T$ 和投影矩阵 $\mathbf{v}_k \mathbf{v}_k^T$ 的乘积为全 0 矩阵。

$$\mathbf{v}_j \mathbf{v}_j^T @ \mathbf{v}_k \mathbf{v}_k^T = \mathbf{O} \quad (48)$$

换个视角来看，图 33 相当于是对图 26 的简化。

图 33. 协方差矩阵可以看成 D 个秩一矩阵取平均

特别地，如果协方差矩阵 Σ 的秩为 r ($r < D$)，则 $\lambda_{r+1}, \dots, \lambda_D$ 均为 0。这种情况下，(47) 可以写成 r 个秩一矩阵相加。

$$\Sigma = \sum_{j=1}^r \lambda_j \mathbf{v}_j \mathbf{v}_j^T + \underbrace{\sum_{j=r+1}^D \lambda_j \mathbf{v}_j \mathbf{v}_j^T}_0 = \sum_{j=1}^r \lambda_j \mathbf{v}_j \mathbf{v}_j^T \quad (49)$$

如图 34 所示, $\Sigma = \mathbf{V} \mathbf{A} \mathbf{V}^T$ 可以写成 $\mathbf{V}^T \Sigma \mathbf{V} = \mathbf{A}$, 展开写成。

$$\begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix} \Sigma [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_D] = \begin{bmatrix} \mathbf{v}_1^T \Sigma \mathbf{v}_1 & \mathbf{v}_1^T \Sigma \mathbf{v}_2 & \cdots & \mathbf{v}_1^T \Sigma \mathbf{v}_D \\ \mathbf{v}_2^T \Sigma \mathbf{v}_1 & \mathbf{v}_2^T \Sigma \mathbf{v}_2 & \cdots & \mathbf{v}_2^T \Sigma \mathbf{v}_D \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_D^T \Sigma \mathbf{v}_1 & \mathbf{v}_D^T \Sigma \mathbf{v}_2 & \cdots & \mathbf{v}_D^T \Sigma \mathbf{v}_D \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix} \quad (50)$$

也就是说 $\mathbf{v}_j^T \Sigma \mathbf{v}_j = \lambda_j$; 当 $j \neq k$ 时, $\mathbf{v}_j^T \Sigma \mathbf{v}_k = 0$ 。

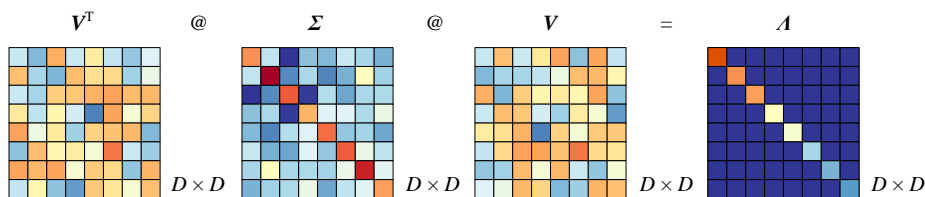


图 34. 把 $\Sigma = \mathbf{V} \mathbf{A} \mathbf{V}^T$ 写成 $\mathbf{V}^T \Sigma \mathbf{V} = \mathbf{A}$

平移 → 旋转 → 缩放

另外, 请大家格外注意多元高斯分布、马氏距离定义蕴含的“平移 → 旋转 → 缩放”, 具体如图 35 所示。

反过来看, 如图 36 所示, 我们也可以通过“缩放 → 旋转 → 平移”将单位球体转化成中心位于任意位置的旋转椭球。

希望这两幅图能够帮助大家回忆仿射变换、椭圆、特征值分解、多元高斯分布、马氏距离、特征值分解、奇异值分解、主成分分析等等书序概念的联系。

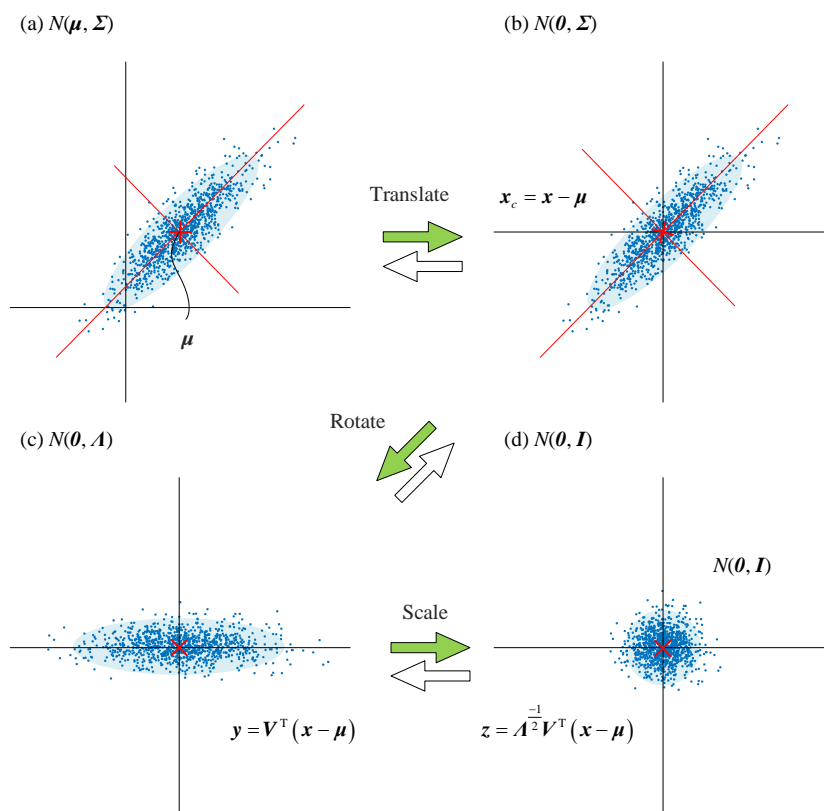


图 35. 旋转椭球，平移 → 旋转 → 缩放

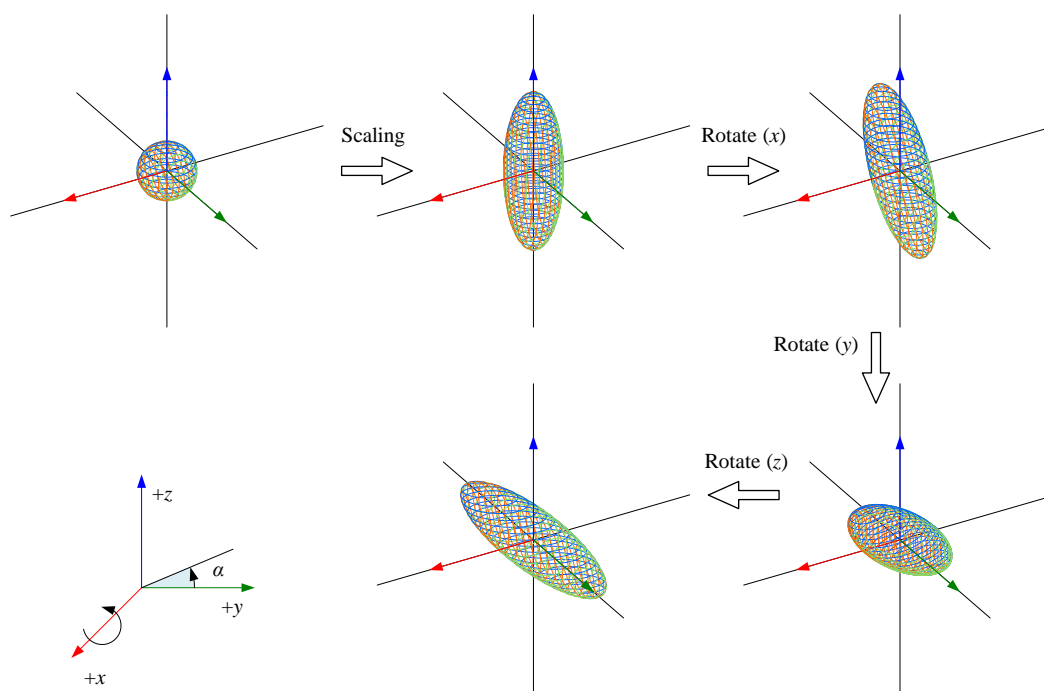


图 36. 旋转椭球，缩放 → 旋转 → 平移

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

线性组合

图 37 所示为原始数据矩阵列向量的线性组合 $y_a = a_1x_1 + a_2x_2 + \dots + a_Dx_D$ ，即

$$y_a = Xa \quad (51)$$

上述线性组合的结果 y_a 是一个列向量，形状为 $n \times 1$ 。

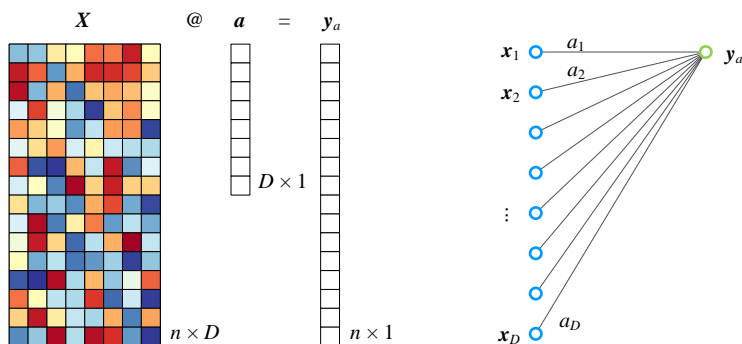


图 37. 原始数据列向量的线性组合

y_a 列向量是一组通过线性组合“人造”的数组，有 n 个样本点。我们很容易计算 y_a 均值。

$$E(y_a) = E(Xa) = E(X)a \quad (52)$$

注意，上式中 $E(X)$ 为行向量，代表数据矩阵 X 的质心。

y_a 的方差。

$$\text{var}(y_a) = \text{var}(Xa) = a^T \Sigma a \quad (53)$$

显然，上式为二次型。鸢尾花书的读者看到“二次型”这三个字，会让我们不禁联想到正定性、EVD、瑞利商、优化问题、标准型、旋转、缩放等等这些数学概念。

如图 38 所示，我们也可以获得原始数据 X 列向量的第二个线性组合，即 $y_b = Xb$ 。我们可以计算 y_b 的均值 $E(y_b)$ 和方差 $\text{var}(y_b)$ ；我们也可以很容易计算得到 y_a 和 y_b 的协方差。

$$\text{cov}(y_a, y_b) = a^T \Sigma b = b^T \Sigma a = \text{cov}(y_b, y_a) \quad (54)$$

其实，(53) 也可以写成 $\text{cov}(y_a, y_a)$ 。

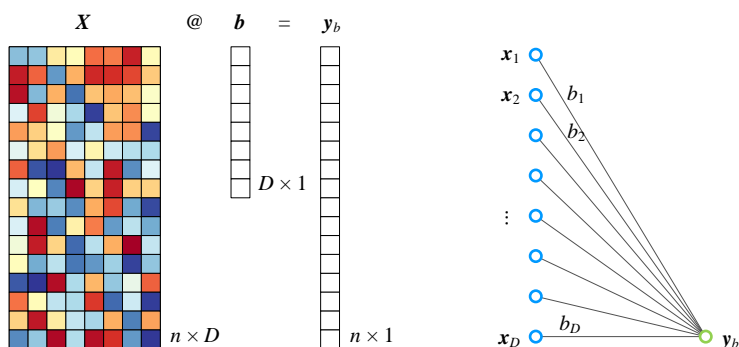


图 38. 原始数据列向量的第二个线性组合

将图 37 和图 38 结合起来，我们便得到了图 39，对应 $Y = XW$ ；也就是 $W = [a, b]$ 。

计算 Y 的协方差矩阵。

$$\text{var}(Y) = \text{var}(XW) = W^T \Sigma W = \begin{bmatrix} a^T \\ b^T \end{bmatrix} \Sigma \begin{bmatrix} a & b \end{bmatrix} = \begin{bmatrix} a^T \Sigma a & a^T \Sigma b \\ b^T \Sigma a & b^T \Sigma b \end{bmatrix} = \begin{bmatrix} \text{var}(y_a) & \text{cov}(y_a, y_b) \\ \text{cov}(y_b, y_a) & \text{var}(y_b) \end{bmatrix} \quad (55)$$

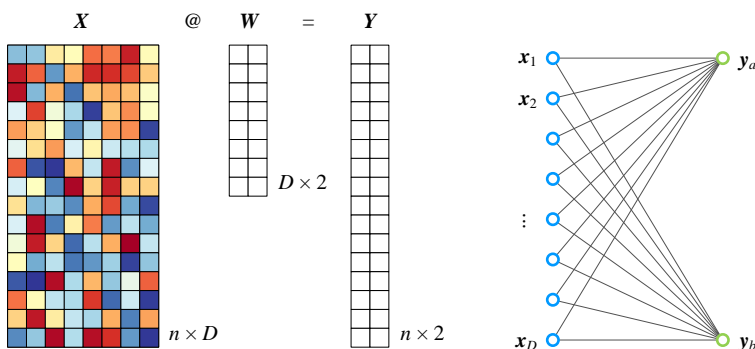


图 39. 原始数据列向量的两个线性组合

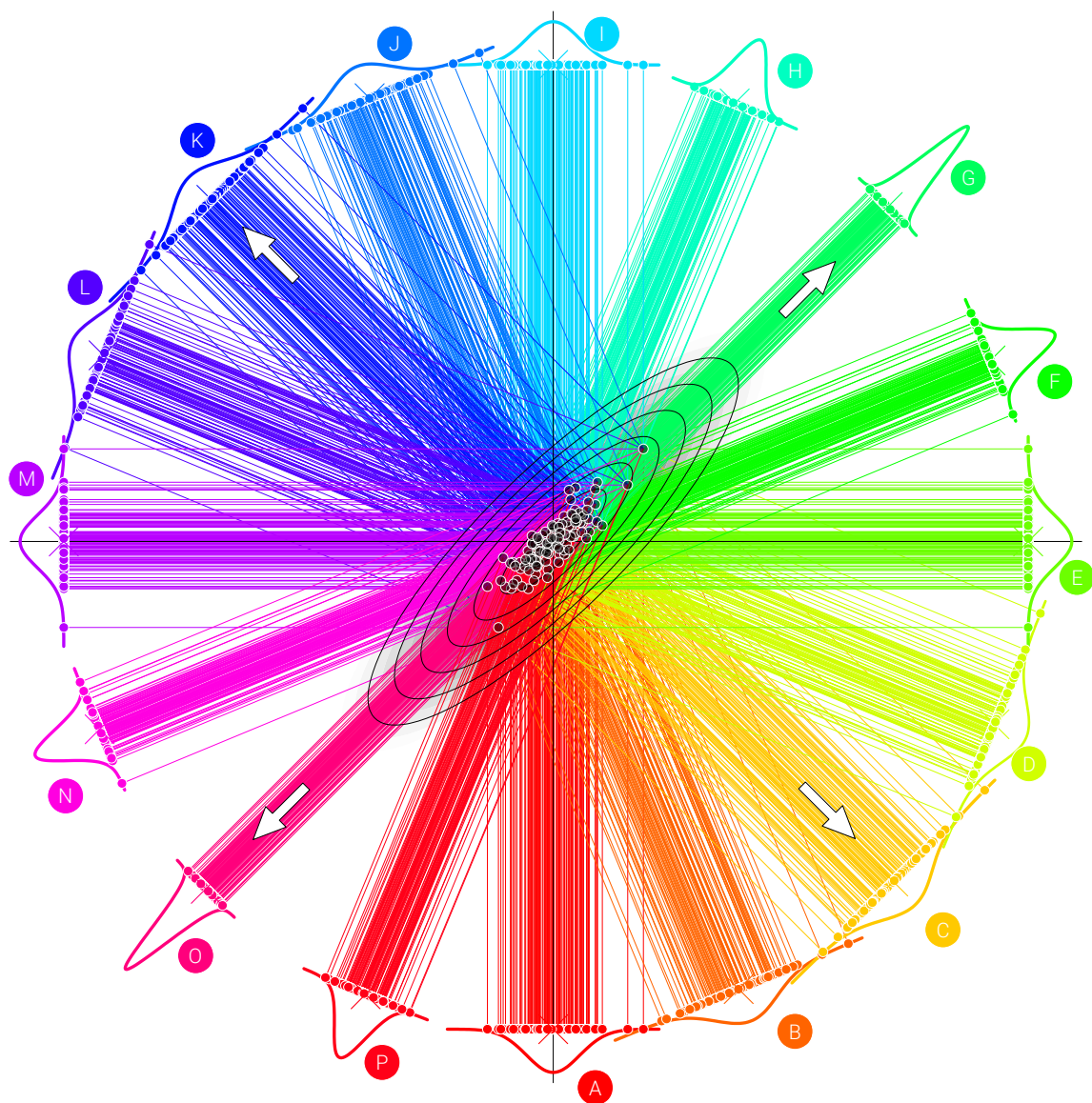
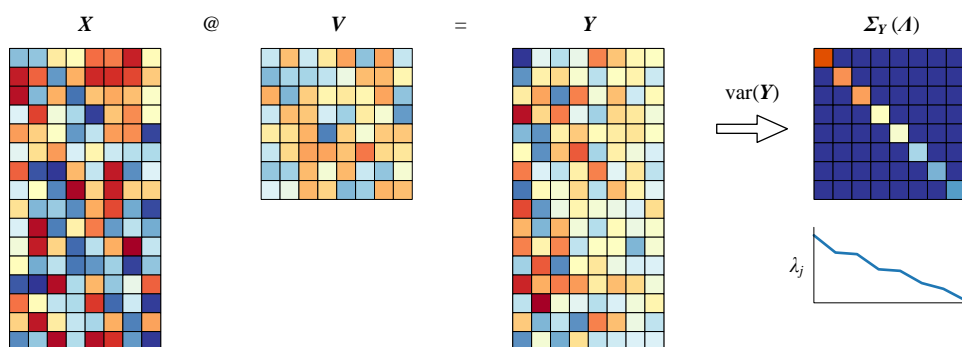
方差最大化

特别地，如果 v 为单位向量，原始数据 X 朝单位向量 v 投影结果为 y ，即 $y = Xv$ 。 y 的方差为。

$$\text{var}(y) = \text{var}(Xv) = v^T \Sigma v \quad (56)$$

如图 40 所示，以二维数据矩阵 X 为例，单位向量 v 不同方向时，我们可以发现 y 的方差有大有小。

而上式的最大值就是协方差矩阵 Σ 的最大特征值 λ_1 ；也就是说， y 的方差最大值为 λ_1 。图 40 这幅图也很好地从几何角度解释了主成分分析。除了特征值分解协方差矩阵，主成分分析还有其他技术路线，这是《机器学习》一册要介绍的内容。

图 40. X 分别朝 16 个不同单位向量投影投影，图片来自《编程不难》图 41. X 投影到 V 空间

如图 41 所示，将数据 X 投影到 V 空间，我们可以得到 Y ，即 $Y = XV$ 。然后，我们可以很容易计算得到 Y 的协方差矩阵

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。

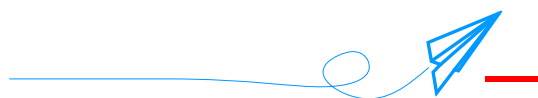
代码及 PDF 文件下载：<https://github.com/Visualize-ML>

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

$$\begin{aligned}\text{var}(\mathbf{Y}) = \text{var}(\mathbf{XV}) &= \mathbf{V}^T \Sigma \mathbf{V} = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_D^T \end{bmatrix} \Sigma [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_D] \\ &= \begin{bmatrix} \mathbf{v}_1^T \Sigma \mathbf{v}_1 & \mathbf{v}_1^T \Sigma \mathbf{v}_2 & \cdots & \mathbf{v}_1^T \Sigma \mathbf{v}_D \\ \mathbf{v}_2^T \Sigma \mathbf{v}_1 & \mathbf{v}_2^T \Sigma \mathbf{v}_2 & \cdots & \mathbf{v}_2^T \Sigma \mathbf{v}_D \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_D^T \Sigma \mathbf{v}_1 & \mathbf{v}_D^T \Sigma \mathbf{v}_2 & \cdots & \mathbf{v}_D^T \Sigma \mathbf{v}_D \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_D \end{bmatrix}\end{aligned}\quad (57)$$

从几何角度来看，这就是“摆正”的椭圆或椭球。



在机器学习中，距离度量是衡量样本之间相似性或差异性的重要指标。在选择距离度量时，需要根据具体问题的性质和数据分布的特点来权衡各种度量的优劣，选择最适合任务的距离度量。

欧氏距离直观且易于理解，计算简单，但是没有考虑特征尺度，也没有考虑数据分布。标准化欧氏距离调整了尺度和单位差异。马氏距离考虑了数据的协方差结构，但是运算成本相对较高。欧氏距离、城市街区距离、切比雪夫距离都是特殊的闵氏距离。

本书后续介绍图论时，大家会看到距离的一种全新形态。

相信有了《矩阵力量》和《统计至简》这两本的铺垫，对于鸢尾花书读者来说，本章有关协方差矩阵的内容应该变得很容易读了。

协方差矩阵是用于衡量多个随机变量之间关系的矩阵。请大家特别注意如何利用椭圆和椭球来理解协方差矩阵。协方差矩阵在机器学习中用途很广，但是协方差矩阵也有自身局限性，请大家注意。

此外，本书第 7 章会介绍用指数加权移动平均计算协方差矩阵；本书第 9 章在讲解高斯过程时，会介绍几种构造先验分布协方差矩阵的核函数。