

11

Matrix Decompositions

矩阵分解

类似代数中的因式分解



宇宙是一部鸿篇巨制，只有掌握它的文字和语言的人才能读懂宇宙；而数学便是解密宇宙的语言。

The universe is a grand book which cannot be read until one first learns to comprehend the language and become familiar with the characters in which it is composed. It is written in the language of mathematics.

—— 伽利略·伽利莱 (Galilei Galileo) | 意大利物理学家、数学家及哲学家 | 1564 ~ 1642



- ◀ matplotlib.pyplot.contour() 绘制等高线图
- ◀ matplotlib.pyplot.contourf() 绘制填充等高线图
- ◀ numpy.linalg.cholesky() Cholesky 分解
- ◀ numpy.linalg.eig() 特征值分解
- ◀ numpy.linalg.qr() QR 分解
- ◀ numpy.linalg.svd() 奇异值分解
- ◀ numpy.meshgrid() 生成网格化数据
- ◀ scipy.linalg.ldl() LDL 分解
- ◀ scipy.linalg.lu() LU 分解
- ◀ seaborn.heatmap() 绘制热图

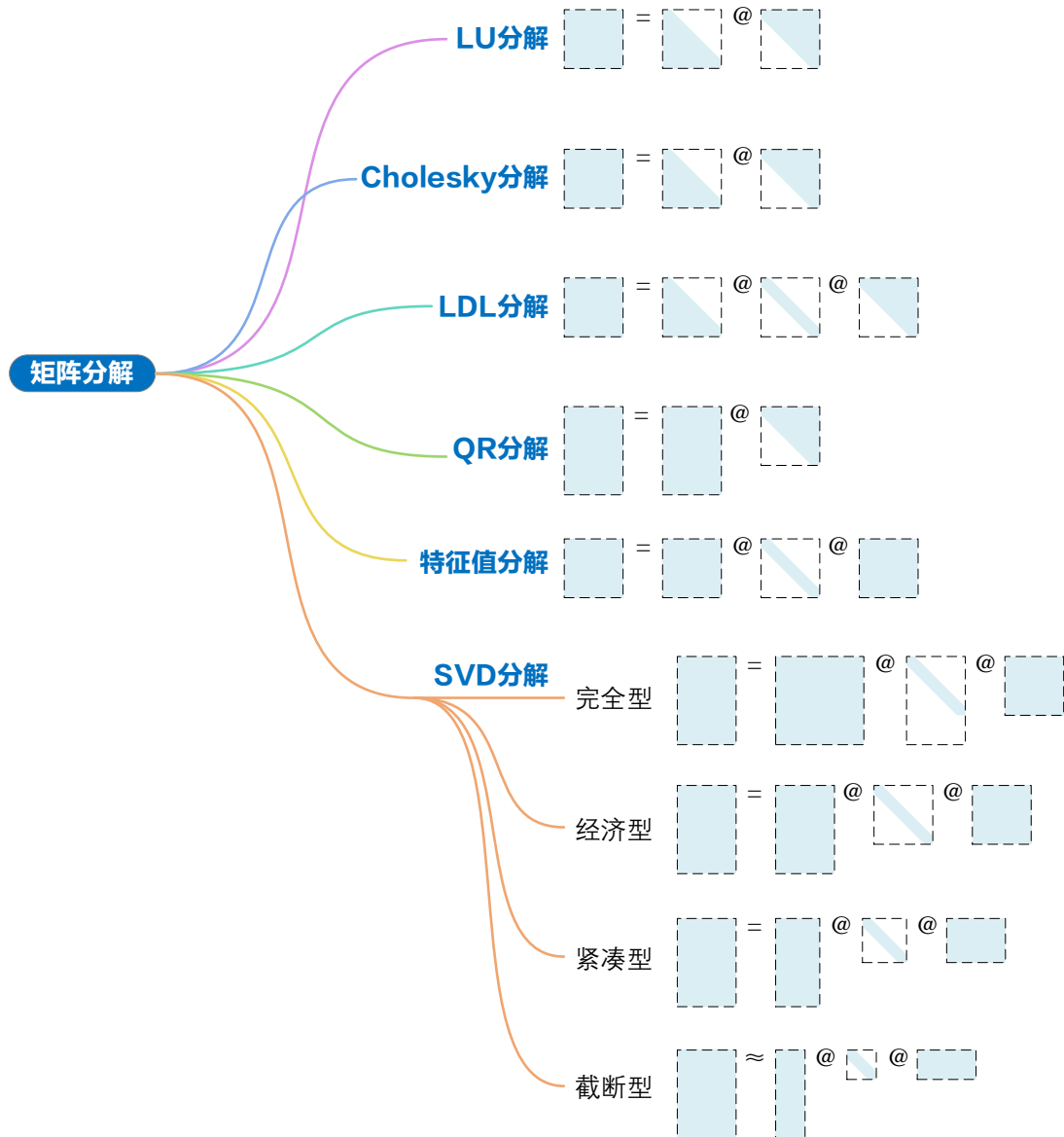
本 PDF 文件为作者草稿，发布目的为方便读者在移动终端学习，终稿内容以清华大学出版社纸质出版物为准。

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



11.1 矩阵分解：类似因式分解

矩阵分解 (matrix decomposition) 将矩阵解构得到其组成部分，类似代数中的因式分解。

从矩阵乘法角度，矩阵分解将矩阵拆解为若干矩阵的乘积。

从几何角度角度，矩阵分解得到的组成部分可能对应缩放、旋转、投影、剪切等等各种几何变换。而原矩阵的映射作用就是这些几何变换按特定次序的叠加。


数据科学和机器学习很多算法都直接依赖矩阵分解。本章全景介绍以下几种矩阵分解：

- ◀ **LU 分解** (lower-upper decomposition, LU decomposition);
- ◀ **Cholesky 分解** (Cholesky decomposition, Cholesky factorization);
- ◀ **LDL 分解** (lower-diagonal-lower transposed decomposition, LDL/LDLT decomposition);
- ◀ **QR 分解** (QR decomposition) 本质上就是本书前文介绍的 Gram-Schmidt 正交化;
- ◀ **特征值分解** (eigendecomposition);
- ◀ **SVD 分解** (singular value decomposition)。

本章偶尔会出现“手算”矩阵分解的情况，这仅仅是为了演示在没有计算机辅助的情况下如何进行特定矩阵分解。本书完全不要求大家掌握矩阵分解“手算”技巧！

此外，仅仅会调用 Numpy 库中函数完成矩阵分解也是远远不够的。

我们需要掌握的是各种不同分解背后的数学思想，更要掌握如何从数据、空间、几何、优化、统计等角度理解这些矩阵分解，并且清楚它们之间的关系、局限性、应用场合。

 在数据分析和机器学习很多算法中，Cholesky 分解、特征值分解和 SVD 分解应用较多，本书此后第 12~16 章将专门讲解这三种矩阵分解。

11.2 LU 分解：上下三角

一说，**LU 分解** (lower-upper decomposition, LU decomposition) 由图灵 (Alan Turing) 于 1948 年发明；另一说，波兰数学家 Tadeusz Banachiewicz 于 1938 年发明 LU 分解。

LU 分解将一个方阵 A ，分解为一个**上三角矩阵** (upper triangular matrix) L 和一个**下三角矩阵** (lower triangular matrix) U 的乘积，即，

$$A = LU \quad (1)$$

(1) 展开得来写：

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,m} \end{bmatrix}_{m \times m} = \begin{bmatrix} l_{1,1} & 0 & \cdots & 0 \\ l_{2,1} & l_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{m,1} & l_{m,2} & \cdots & l_{m,m} \end{bmatrix}_{m \times m} \begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,m} \\ 0 & u_{2,2} & \cdots & u_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{m,m} \end{bmatrix}_{m \times m} \quad (2)$$

图 1 所示为 LU 分解对应的矩阵运算示意图。LU 分解可以视为**高斯消元法** (Gaussian elimination) 的矩阵乘法形式。

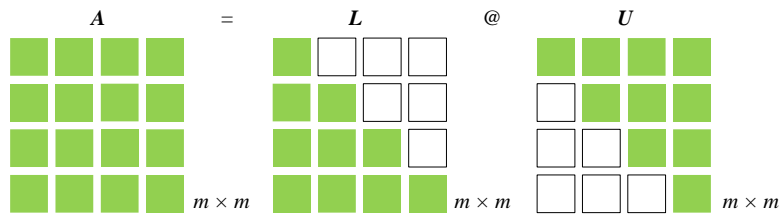


图 1. LU 分解

本书常用 `scipy.linalg.lu()` 函数进行 LU 分解。注意, `scipy.linalg.lu()` 默认进行 PLU 分解, 即,

$$A = PLU \quad (3)$$

其中, **P** 为**置换矩阵** (permutation matrix)。`scipy.linalg.lu()` 函数得到的矩阵 **L** 主对角线均为 1。

前文介绍过, 置换矩阵的任意一行或列只有一个 1, 剩余均为 0。置换矩阵的作用是交换矩阵的行、列。

▲ 注意, 本书中默认置换矩阵为方阵。置换矩阵的逆还是置换矩阵, 置换矩阵必定是正交矩阵。

图 2 所示为对方阵 **A** 进行 PLU 分解运算热图。注意, 所有的方阵都可以进行 PLU 分解。

PLU 分解有很高的数值稳定性。举个例子, 如果 (1) 中矩阵 **A** 中有一个元素的数值特别小, LU 分解后, 得到的 **L** 和 **U** 矩阵会出现数值很大的数。为了避免这种情况, 如 (3) 所示, 通过一个置换矩阵 **P**, 先对矩阵 **A** 进行变换, 然后再进行 LU 分解。

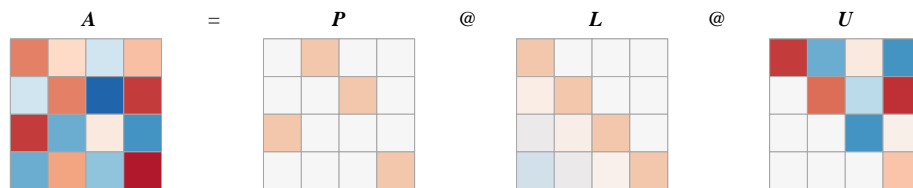
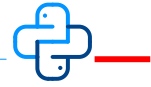


图 2. 对矩阵 **A** 的 PLU 分解热图



Bk4_Ch11_01.py 绘制图 2。

11.3 Cholesky 分解：得到数据矩阵的坐标

Cholesky 分解 (Cholesky decomposition) 是 LU 分解的特例。丛书在讲解**协方差矩阵** (covariance matrix)、数据转换、蒙特卡洛模拟等内容都会使用 Cholesky 分解。

Cholesky 分解把矩阵分解为一个下三角矩阵以及它的转置矩阵的乘积：

$$A = LL^T \quad (4)$$

也就是说：

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,m} \end{bmatrix}_{m \times m} = \begin{bmatrix} l_{1,1} & 0 & \cdots & 0 \\ l_{2,1} & l_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{m,1} & l_{m,2} & \cdots & l_{m,m} \end{bmatrix}_{m \times m} \begin{bmatrix} l_{1,1} & l_{2,1} & \cdots & l_{m,1} \\ 0 & l_{2,2} & \cdots & l_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_{m,m} \end{bmatrix}_{m \times m} \quad (5)$$

当然，利用上三角矩阵 R ，Cholesky 分解也可以写成：

$$A = R^T R \quad (6)$$

其中， $R = L^T$ 。

图 3 所示为 Cholesky 分解矩阵运算示意图。本节标题说 Cholesky 分解“得到数据矩阵的坐标”，这个话题留给下一章。

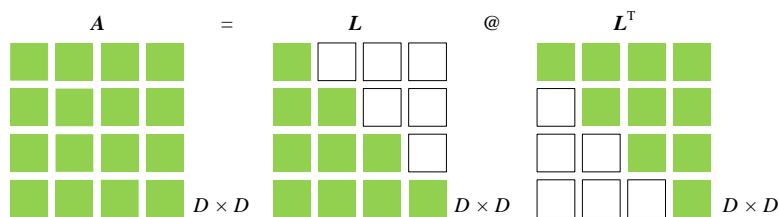


图 3. Cholesky 分解矩阵运算

Numpy 中进行 Cholesky 分解的函数为 `numpy.linalg.cholesky()`。请读者自行编写代码并绘制图 4。

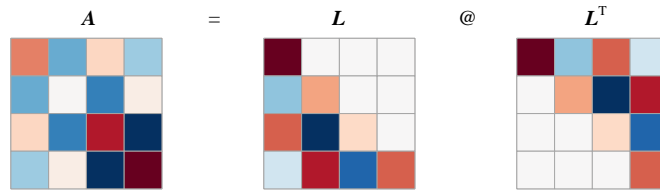


图 4. Cholesky 分解示例

▲ 注意，只有正定矩阵 (positive definite matrix) 才能 Cholesky 分解。下一章将简单地介绍正定性及其几何内涵。本书第 21 章将专门讲解正定性。

LDL 分解：Cholesky 分解的扩展

Cholesky 分解可以进一步扩展为 **LDL 分解** (LDL decomposition):

$$A = LDL^T = LD^{1/2} (D^{1/2})^T L^T = LD^{1/2} (LD^{1/2})^T \quad (7)$$

其中， L 为下三角矩阵，但是对角线元素均为 1； D 为对角矩阵，起到缩放作用。

(7) 展开来写得到：

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,m} \end{bmatrix}_{m \times m} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ l_{2,1} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{m,1} & l_{m,2} & \cdots & 1 \end{bmatrix}_{m \times m} \begin{bmatrix} d_{1,1} & 0 & \cdots & 0 \\ 0 & d_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_{m,m} \end{bmatrix}_{m \times m} \begin{bmatrix} 1 & l_{2,1} & \cdots & l_{m,1} \\ 0 & 1 & \cdots & l_{m,2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{m \times m} \quad (8)$$

图 5 所示为 LDL 分解矩阵运算示意图。

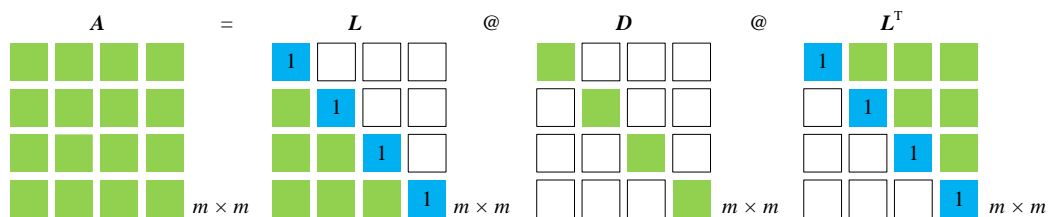


图 5. LDL 分解矩阵运算示意图

LDL 分解的函数为 `scipy.linalg.ldl()`，注意这个函数的返回结果也包括置换矩阵。图 6 所示为 LDL 分解运算热图。请读者根据前文代码自行绘制图 6。

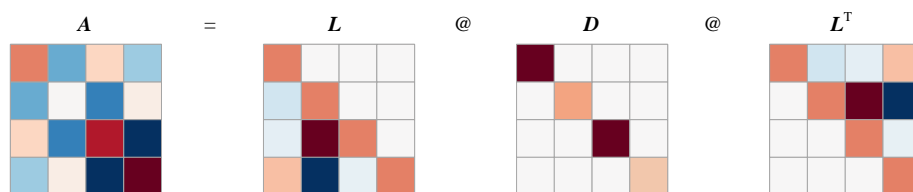


图 6. LDL 分解示例热图

11.4 QR 分解：正交化

QR 分解 (QR decomposition, QR factorization) 和本书第 9 章介绍的格拉姆-斯密特正交化联系紧密。QR 分解有两种常见形式：

- ◀ **完全型** (complete), Q 为方阵；
- ◀ **缩略型** (reduced), Q 和原矩阵形状相同。

图 7 所示为形状对 $n \times D$ 数据矩阵 X 进行完全型 QR 分解示意图，对应的等式为：

$$X_{n \times D} = Q_{n \times n} R_{n \times D} \quad (9)$$

其中， Q 为方阵，形状为 $n \times n$ ； R 和 X 形状一致，形状为 $n \times D$ 。

⚠ 注意，任何实数矩阵都可以 QR 分解。

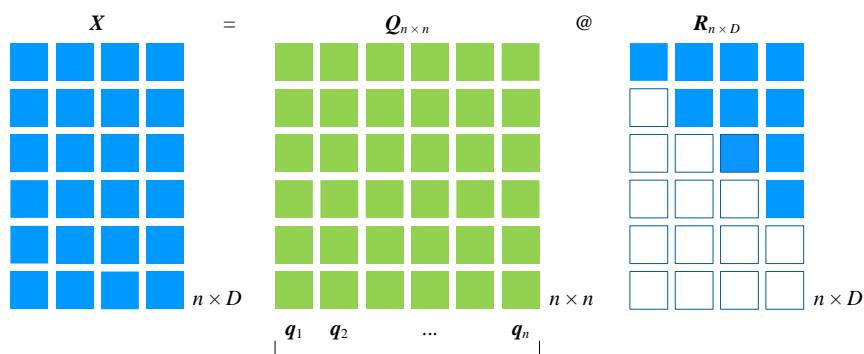


图 7. 完全型 QR 分解示意图

图 8 所示为对某个细高数据矩阵 X 进行完全型 QR 分解运算热图。

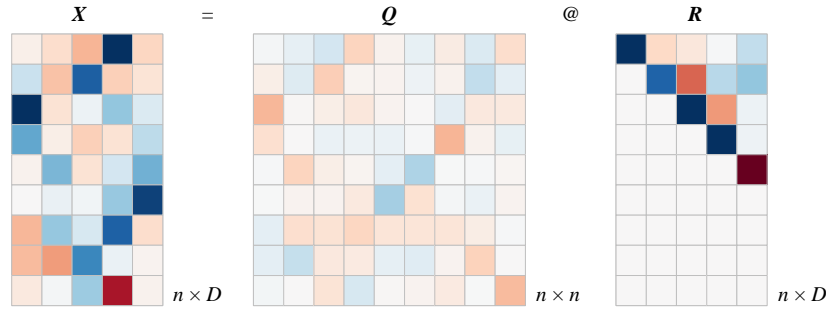
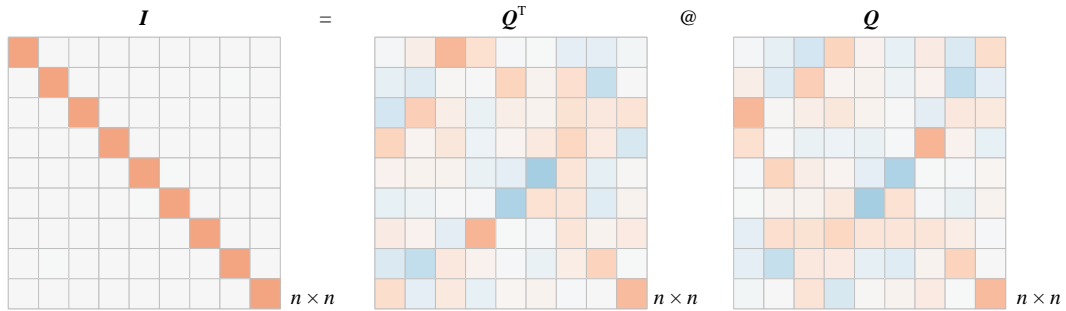


图 8. 完全型 QR 分解热图

方阵 Q 为正交矩阵，也就是说：

$$Q_{n \times n} Q_{n \times n}^T = Q_{n \times n}^T Q_{n \times n} = I_{n \times n} \quad (10)$$

图 9 所示为 (10) 运算对应热图。根据本书前文介绍的有关正交矩阵的性质， $Q = [q_1, q_2, \dots, q_D]$ 是一个规范正交基，张起的向量空间为 \mathbb{R}^D 。

图 9. Q_c 为正交矩阵

把 Q 展开写成 $[q_1, q_2, \dots, q_D]$ ， X 的第一列向量 x_1 可以通过下式得到：

$$x_1 = [q_1 \quad q_2 \quad \dots \quad q_D] \begin{bmatrix} r_{1,1} \\ r_{2,1} \\ \vdots \\ r_{D,1} \end{bmatrix} = r_{1,1} q_1 + \underbrace{r_{2,1} q_2}_{=0} + \dots + \underbrace{r_{D,1} q_D}_{=0} = r_{1,1} q_1 \quad (11)$$

上式相当于 x_1 在规范正交基 $[q_1, q_2, \dots, q_D]$ 张成的空间坐标为 $(r_{1,1}, r_{2,1}, \dots, r_{D,1})$ ，即 $(r_{1,1}, 0, \dots, 0)$ 。也就是说， x_1 和 q_1 平行，方向同向或反向。这和本书第 9 章介绍的格拉姆-施密特正交化第一步一致。

q_1 是单位向量，也就是说：

$$r_{1,1} = \pm \|x_1\| \quad (12)$$

这一点已经说明 QR 分解结果不唯一。但是，如果 X 列满秩，且 R 的对角元素为正实数的情况下 QR 分解唯一。

类似地， X 的第二列向量 x_2 写成：

$$x_2 = [q_1 \quad q_2 \quad \cdots \quad q_D] \begin{bmatrix} r_{1,2} \\ r_{2,2} \\ \vdots \\ r_{D,2} \end{bmatrix} = r_{1,2}q_1 + r_{2,2}q_2 + \underbrace{r_{3,2}q_3}_{=0} + \cdots + \underbrace{r_{D,2}q_D}_{=0} = r_{1,2}q_1 + r_{2,2}q_2 \quad (13)$$

x_2 在规范正交基 $[q_1, q_2, \dots, q_D]$ 张成的空间坐标为 $(r_{1,2}, r_{2,2}, r_{3,2}, \dots, r_{D,2})$ ，即 $(r_{1,2}, r_{2,2}, 0, \dots, 0)$ 。

缩略型

图 7 对应的完全型 QR 分解可以进一步简化。将 (9) 中 R 上下切一刀，让上方子块为方阵，下方子块为零矩阵 O 。这样 (9) 可以写成分块矩阵乘法：

$$X = \begin{bmatrix} Q_{n \times D} & Q_{n \times (n-D)} \end{bmatrix} \begin{bmatrix} R_{D \times D} \\ O_{(n-D) \times D} \end{bmatrix} = Q_{n \times D} R_{D \times D} + Q_{n \times (n-D)} O_{(n-D) \times D} = Q_{n \times D} R_{D \times D} \quad (14)$$

其中， $Q_{n \times D}$ 和 X 矩阵形状相同，而 $R_{D \times D}$ 为上三角方阵。注意，上式中零矩阵 O 的形状为 $(n-D) \times D$ ，其所有元素均为 0。

图 10 所示为 QR 分解从完全型到缩略型简化过程。

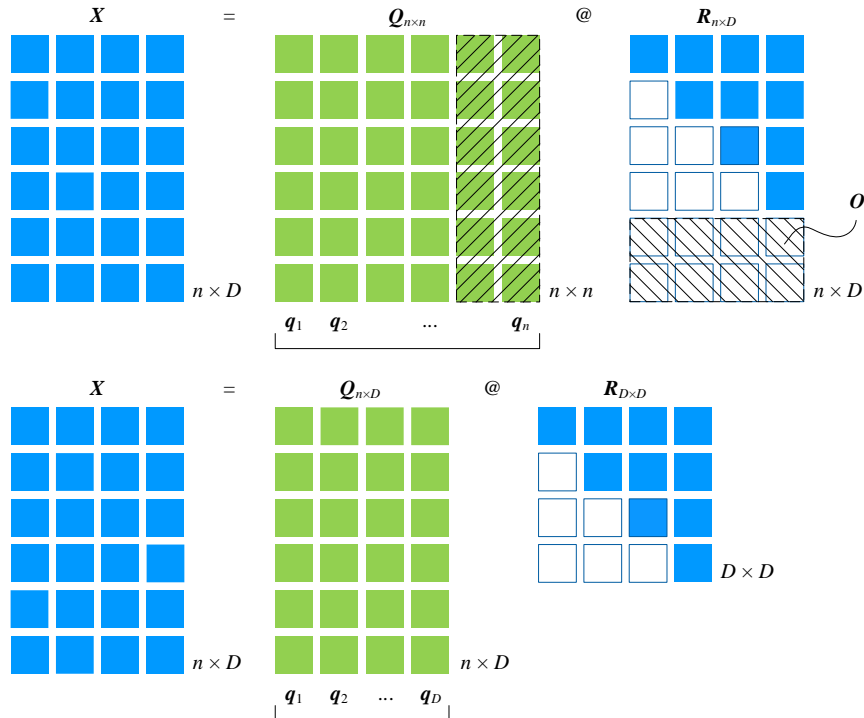


图 10. QR 分解从完全型到缩略型简化过程

图 11 所示为对矩阵 X 进行缩略型 QR 分解运算热图。

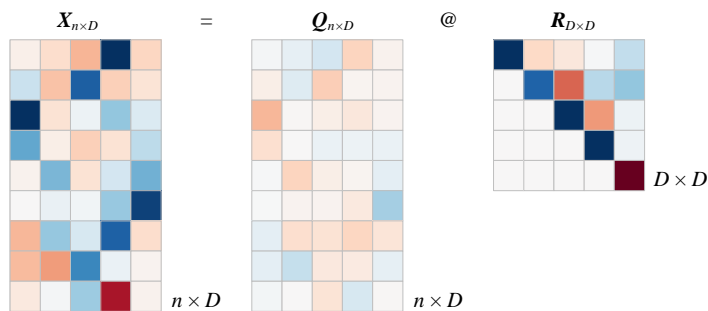


图 11. 缩略型 QR 分解热图

列向量两两正交

虽然 (14) 中矩阵 $Q_{n \times D}$ 不是一个方阵，但列向量也两两正交，因为，

$$(Q_{n \times D})^T Q_{n \times D} = I_{D \times D} \quad (15)$$

把 Q 展开写成 $[q_1, q_2, \dots, q_D]$ ，代入上式得到：

$$\begin{aligned} Q^T Q &= \begin{bmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_D^T \end{bmatrix} [q_1 \ q_2 \ \cdots \ q_D] \\ &= \begin{bmatrix} q_1^T q_1 & q_1^T q_2 & \cdots & q_1^T q_D \\ q_2^T q_1 & q_2^T q_2 & \cdots & q_2^T q_D \\ \vdots & \vdots & \ddots & \vdots \\ q_D^T q_1 & q_D^T q_2 & \cdots & q_D^T q_D \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{D \times D} \end{aligned} \quad (16)$$

其中， q_j 向量为 n 行。图 12 所示为 $Q^T Q$ 运算对应的热图。

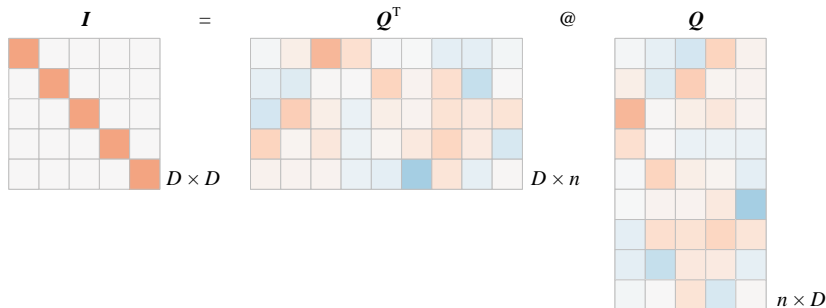


图 12. $Q^T Q$ 运算对应的热图

几何视角

QR 分解完成对数据矩阵 X 的正交化。 X 的列向量 $[x_1, x_2, \dots, x_D]$ 可能并非两两正交，经过 QR 分解得到的 $[q_1, q_2, \dots, q_D]$ 两两正交，且每个向量为单位向量。如图 13 所示， $[q_1, q_2, \dots, q_D]$ 本身就是一个规范正交基，它的重要特点是 q_1 平行于 x_1 ，通过逐步正交投影得到剩余 $q_j (j = 2, 3, \dots, D)$ 。

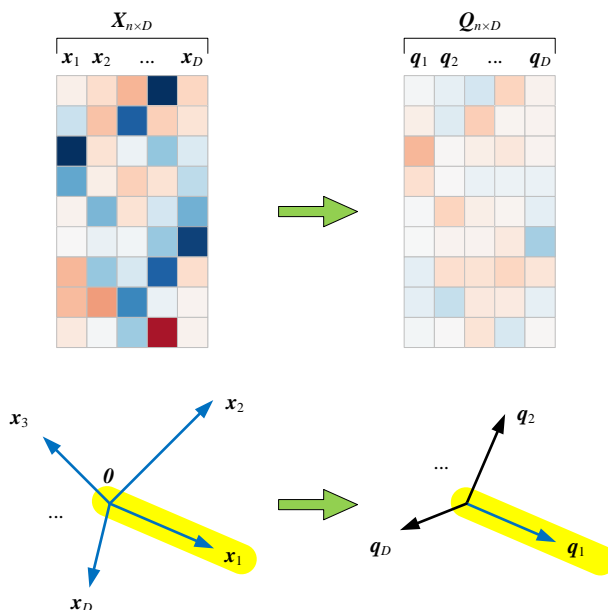


图 13. QR 分解背后的几何意义

当然，对数据矩阵 X 的正交化方法并不唯一，不同正交化方法得到的规范正交基也不同。本书后面还会介绍其他正交化方法，请大家注意区分结果的差异以及应用场合。



Bk4_Ch11_02.py 绘制本节热图。

11.5 特征值分解：刻画矩阵映射的特征

枯燥的定义

对于方阵 A ，如果存在**非零向量** (non-zero vector) \mathbf{v} 使得：

$$A\mathbf{v} = \lambda\mathbf{v} \quad (17)$$

\mathbf{v} 就是的 A 的**特征向量** (eigen vector)，标量 λ 被称作**特征值** (eigen value)。

(17) 可以写作：

$$(A - \lambda I)\mathbf{v} = 0 \quad (18)$$

其中， I 是**单位阵** (identity matrix)。

特征向量 \mathbf{v} 代表方向，通常是列向量；特征值 λ 是在这个方向上的比例，特征值是标量。

⚠ 注意，并不是所有方阵都可以特征值分解，只有可对角化矩阵 (diagonalizable matrix) 才能进行特征值分解。如果一个方阵 A 相似于对角矩阵，也就是说，如果存在一个可逆矩阵 V 使得矩阵乘积 $V^{-1}AV$ 结果为对角矩阵，则 A 就被称为可对角化 (diagonalizable)。大家是否还记得，本书前文讲解几何变换时提到，几何上，我们更喜欢看到对角阵，因为在空间中对角阵代表“立方体”。

二维方阵

假设某个二维方阵 A ，有两个特征值和特征向量：

$$\begin{aligned} A\mathbf{v}_1 &= \lambda_1\mathbf{v}_1 \\ A\mathbf{v}_2 &= \lambda_2\mathbf{v}_2 \end{aligned} \quad (19)$$

两个特征向量可以构成矩阵 V ，用两个特征值构造对角阵 A ，上式可以写成：

$$A \underbrace{\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix}}_{\mathbf{V}} = \underbrace{\begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 \end{bmatrix}}_{\mathbf{V}} \underbrace{\begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}}_{\mathbf{A}} \quad (20)$$

即，

$$AV = VA \quad (21)$$

(21) 可以进一步写成：

$$A = VAV^{-1} \quad (22)$$

上式就叫做矩阵 A 的**特征分解** (eigen-decomposition)。 A 被称作特征值矩阵， V 被称作特征向量矩阵。

多维方阵

对于 $D \times D$ 方阵 A ，如果存在如下一系列等式：

$$\begin{cases} \mathbf{A}\mathbf{v}_1 = \lambda_1\mathbf{v}_1 \\ \mathbf{A}\mathbf{v}_2 = \lambda_2\mathbf{v}_2 \\ \vdots \\ \mathbf{A}\mathbf{v}_D = \lambda_D\mathbf{v}_D \end{cases} \quad (23)$$

整理上式得到：

$$[\mathbf{A}\mathbf{v}_1 \quad \mathbf{A}\mathbf{v}_2 \quad \cdots \quad \mathbf{A}\mathbf{v}_D] = [\lambda_1\mathbf{v}_1 \quad \lambda_2\mathbf{v}_2 \quad \cdots \quad \lambda_D\mathbf{v}_D] \quad (24)$$

即，

$$\mathbf{A} \underbrace{[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_D]}_{\mathbf{V}} = \underbrace{[\lambda_1\mathbf{v}_1 \quad \lambda_2\mathbf{v}_2 \quad \cdots \quad \lambda_D\mathbf{v}_D]}_{\mathbf{V}\mathbf{\Lambda}} \quad (25)$$

特征多项式

方阵 \mathbf{A} **特征多项式** (characteristic polynomial) 可以这样获得：

$$p(\lambda) = |\mathbf{A} - \lambda\mathbf{I}| \quad (26)$$

\mathbf{A} 的**特征方程** (characteristic equation) 为：

$$|\mathbf{A} - \lambda\mathbf{I}| = 0 \quad (27)$$

特征方程可以用来求解矩阵的特征值，从而进一步求解对应的特征向量。

手算特征值分解

给定如下矩阵 \mathbf{A} ：

$$\mathbf{A} = \begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{bmatrix} \quad (28)$$

方阵 \mathbf{A} 的特征方程为：

$$\begin{aligned} p(\lambda) = |\mathbf{A} - \lambda\mathbf{I}| &= \begin{vmatrix} 1.25 - \lambda & -0.75 \\ -0.75 & 1.25 - \lambda \end{vmatrix} \\ &= \lambda^2 - 2.5\lambda + 1 = (\lambda - 2)(\lambda - 0.5) = 0 \end{aligned} \quad (29)$$

求解 (29) 所示一元二次方程，得到 $p(\lambda)$ 的两个根分别为：

$$\lambda_1 = 0.5, \quad \lambda_2 = 2 \quad (30)$$

对于 $\lambda_1 = 0.5$,

$$(A - \lambda_1 I)v_1 = \left\{ \begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{bmatrix} - \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \right\} \begin{bmatrix} v_{1,1} \\ v_{1,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (31)$$

得到如下等式：

$$v_{1,1} - v_{1,2} = 0 \quad (32)$$

满足如上等式的向量都是特征向量，选择第一象限的单位向量为特征向量 v_1

$$v_1 = \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix} \quad (33)$$

这一步可以看出特征向量不唯一。本书中，特征向量一般都是单位向量，除非特殊说明。

对于 $\lambda_2 = 2$,

$$(A - \lambda_2 I)v_2 = \left\{ \begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{bmatrix} - \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \right\} \begin{bmatrix} v_{2,1} \\ v_{2,2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (34)$$

得到如下等式：

$$v_{2,1} + v_{2,2} = 0 \quad (35)$$

同样，满足如上等式的向量都是特征向量，选择第二象限的单位向量为特征向量 v_2

$$v_2 = \begin{bmatrix} -\sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix} \quad (36)$$

图 14 所示为候选特征向量之间关系。

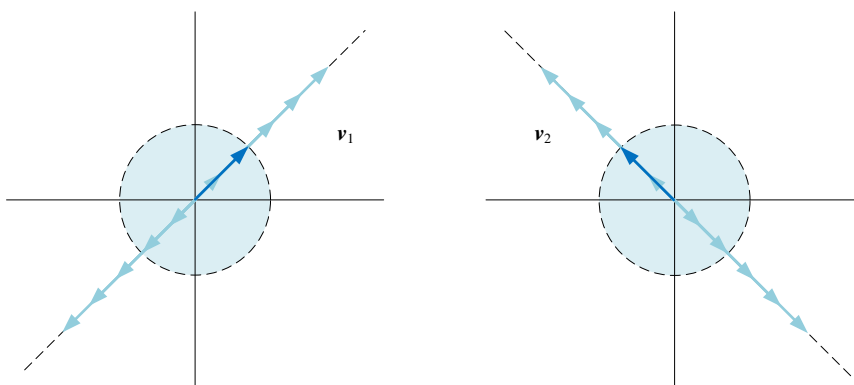


图 14. 候选特征向量

这样我们可以得到特征向量矩阵 V ：

$$\mathbf{V} = \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \quad (37)$$

\mathbf{V} 的逆为：

$$\mathbf{V}^{-1} = \begin{bmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} \quad (38)$$

大家可能已经发现：

$$\mathbf{V}^T = \mathbf{V}^{-1} \quad (39)$$

这是因为 (28) 中 \mathbf{A} 为对称矩阵。这是特征值分解的一种特殊情况。

对称矩阵

如果 \mathbf{A} 为对称矩阵，则 (22) 可以写作：

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \quad (40)$$

\mathbf{V} 为正交矩阵，即满足：

$$\mathbf{V}\mathbf{V}^T = \mathbf{I} \quad (41)$$

几何视角

对于一个细高的长方形矩阵 \mathbf{X} 来说，它本身肯定不能进行特征值分解。但是，它的两个格拉姆矩阵 $\mathbf{X}^T\mathbf{X}$ 和 $\mathbf{X}\mathbf{X}^T$ 都是对称阵！强调一下，这里 \mathbf{X} 为实数矩阵， $\mathbf{X}^T\mathbf{X}$ 和 $\mathbf{X}\mathbf{X}^T$ 都是对称矩阵，因此可以对角化。

如图 15 所示， $\mathbf{X}^T\mathbf{X}$ 和 $\mathbf{X}\mathbf{X}^T$ 都可以进行特征值分解，而且分解得到的特征向量矩阵 \mathbf{V} 和 \mathbf{U} 都是正交矩阵。 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_D]$ 张起的向量空间为 \mathbb{R}^D 。 $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n]$ 张起的向量空间为 \mathbb{R}^n 。

之所以用 \mathbf{V} 和 \mathbf{U} 分别表达特征向量矩阵，是为了和下一节奇异值分解呼应。

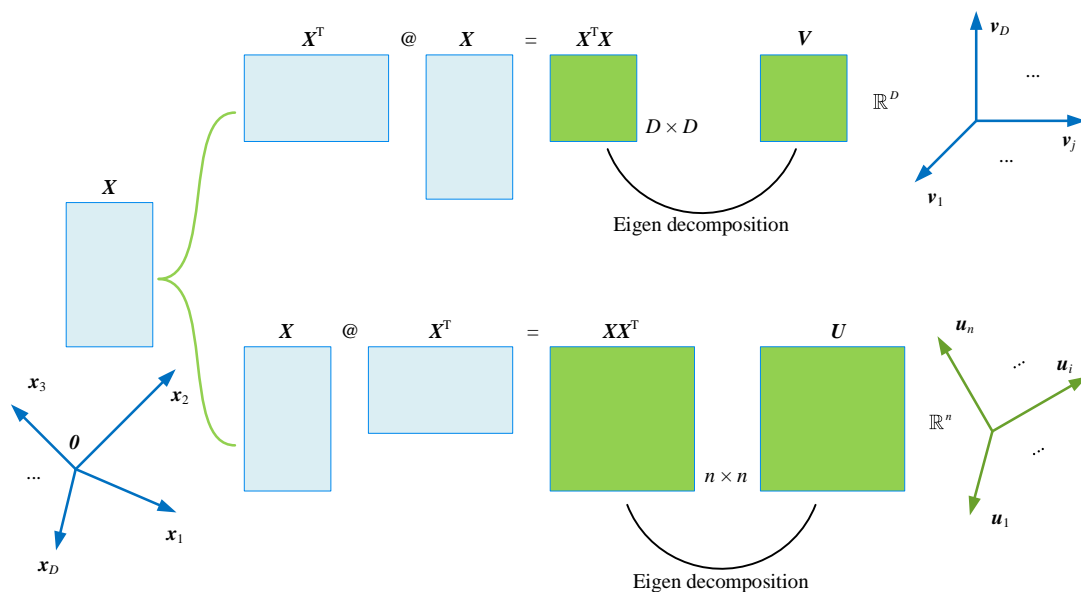


图 15. 对 Gram 矩阵特征值分解

➡ 如果本书有关特征值分解内容就此结束的话，相信所有读者会极其失望，说好的“图解”呢？空间、几何、数据、优化、统计视角又在哪？特征值分解是矩阵分解中的一道“大菜”，它在数据科学和机器学习领域应用非常广泛，本节仅仅介绍其皮毛。本书第 13、14 章专门讲解特征值分解及其应用。

11.6 奇异值分解：适用于任何实数矩阵

如果特征值分解是“大菜”，奇异值分解绝对就是矩阵分解中的“头牌”！本节将蜻蜓点水地介绍一些奇异值分解最基本概念，并让大家尝尝手算奇异值分解的滋味！

➡ 本书第 15、16 两章专门讲解奇异值分解和应用。本书最后三章还会梳理特征值分解和奇异值分解之间关系，以及它们和数据、空间、统计等概念的关系，把大家对矩阵分解的认识提高一个全新高度。

定义

对矩阵 $X_{n \times D}$ **奇异值分解** (Singular Value Decomposition, SVD)，得到：

$$X_{n \times D} = USV^T \quad (42)$$

S 主对角线元素 s_i 为**奇异值** (singular value)。一些教材用 Σ 代表奇异值矩阵，而丛书专门用 Σ 作为协方差矩阵记号。本书偶尔会用 S 代表“缩放”矩阵，这和奇异值分解中的 S 在功能上完全一致。

U 的列向量称作**左奇异值向量** (left singular vector)。

V 的列向量称作**右奇异值向量** (right singular vector)。

常用的 SVD 分解有四种类型。完全型 SVD 分解中， U 和 V 为方阵， S 和 X 的形状相同，具体如图 16 所示。本书第 15、16 章会介绍 SVD 的四种分解类型。

⚠ 注意，任何实数矩阵都可以 SVD 分解。“任何”二字奠定了奇异值分解宇宙第一矩阵分解的地位！不管是方阵，还是细高、矮胖长方阵，SVD 分解都能处理，可谓兵来将挡、水来土掩。

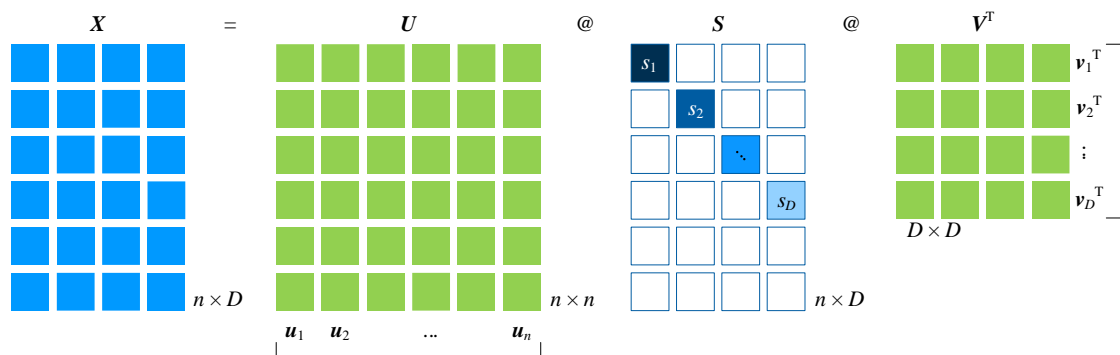


图 16. SVD 分解示意图

两个规范正交基

如果 U 和 V 都是方阵的话，它们都是正交矩阵。这也就是说，几何视角下， U 和 V 都是规范正交基！如图 17 所示，这相当于一个 SVD 完成了图 15 中两个特征值分解。

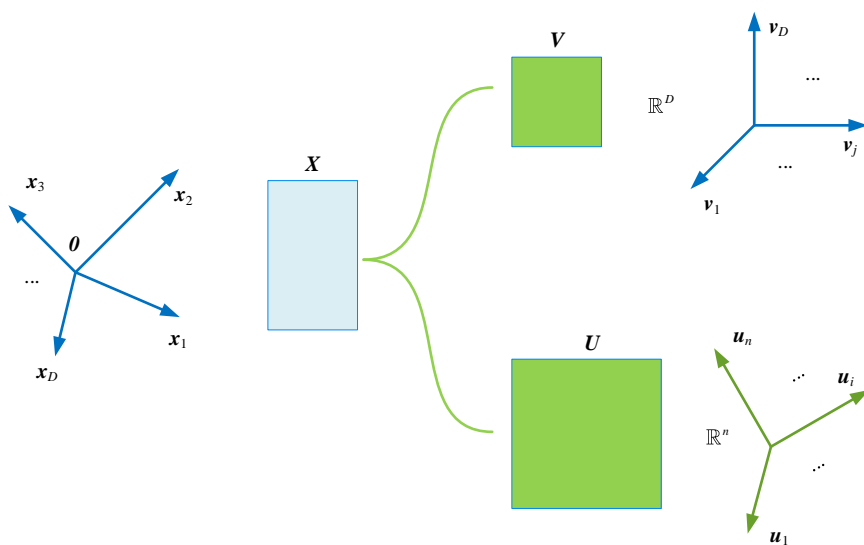


图 17. 对 X 矩阵 SVD 分解获得两个规范正交基



SVD 分解也是对原始数据矩阵进行正交化的工具，本章前文提到 QR 分解和特征值分解都可以得到规范正交基，这些矩阵分解之间的区别和联系是什么？得到的规范正交基有什么不同？它们和向量空间又有怎样关系？这是本书最后三章“数据三部曲”要回答的问题。

手算奇异值分解

给定矩阵 X :

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \quad (43)$$

为求解 V ，先计算第一个格拉姆矩阵—— $X^T X$,

$$X^T X = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}^T \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad (44)$$

前文提过，细高或矮胖的长方形矩阵在进行矩阵运算时并不友好，我们通常需要将它们“平方”，写成格拉姆矩阵 $X^T X$ 这种形式，以便完成进一步运算。此外，本书前文还提到 Gram 矩阵的几何内涵，下一章大家可以看到 Cholesky 分解和格拉姆矩阵之间碰撞出的火花。

进一步计算得到 $X^T X$ 特征值和特征向量：

$$\left\{ \begin{array}{l} \lambda_1 = 3 \\ v_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \end{array} \right\} \quad \left\{ \begin{array}{l} \lambda_2 = 1 \\ v_2 = \begin{bmatrix} -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix} \end{array} \right\} \quad (45)$$

然后，计算第二个格拉姆矩阵—— XX^T ,

$$XX^T = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}^T = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad (46)$$

⚠ 注意区分， $X^T X$ 形状为 2×2 ， XX^T 形状为 3×3 。两者都是 Gram 矩阵，前者关注 X 的列向量，后者关注 X 的行向量。

计算 XX^T 特征值和特征向量：

$$\left\{ \begin{array}{l} \lambda_1 = 3 \\ \mathbf{u}_1 = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} \end{array} \right\} \left\{ \begin{array}{l} \lambda_2 = 1 \\ \mathbf{u}_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ -\frac{\sqrt{2}}{2} \end{bmatrix} \end{array} \right\} \left\{ \begin{array}{l} \lambda_3 = 0 \\ \mathbf{u}_3 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{3}}{3} \\ \frac{\sqrt{3}}{3} \end{bmatrix} \end{array} \right\} \quad (47)$$

奇异值矩阵如下：

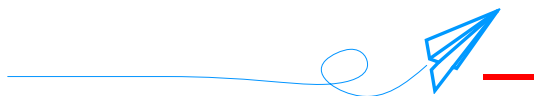
$$\mathbf{S} = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \quad (48)$$

⚠ 注意，对角线元素为 λ_1 和 λ_2 平方根。这一点是特征值分解和 SVD 分解的一个重要的区别，也是一个重要的联系。请大家格外留意。

因此，对 \mathbf{X} 进行完全型 SVD 分解可以得到：

$$\mathbf{X} = \mathbf{USV}^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \\ \frac{2}{\sqrt{6}} & 0 & -\frac{\sqrt{3}}{3} \\ \frac{1}{\sqrt{6}} & -\frac{\sqrt{2}}{2} & \frac{\sqrt{3}}{3} \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}^T \quad (49)$$

再次强调，本书绝不要求大家掌握如何徒手进行 SVD 分解。需要大家掌握的是 SVD 背后的数学思想，如何利用不同视角理解 SVD 分解以及它的应用。



本章开启了本书一个全新的板块——矩阵分解。以下四幅图总结本章的主要内容。

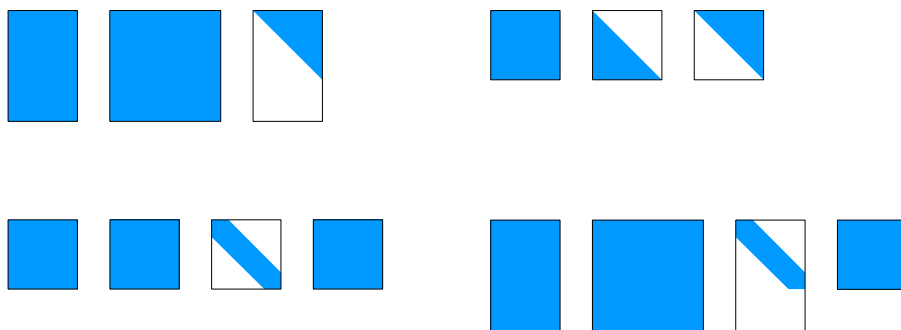


图 18. 总结本章重要内容的四副图

矩阵分解看着让人眼花缭乱，但是万变不离其宗，矩阵分解的内核还是矩阵乘法！大家很快就会看到，我们会反复使用矩阵乘法的两个视角来分析各种矩阵分解。

矩阵分解让我们从一个全新的高度领略到了矩阵乘法的魅力。数据视角、几何视角，这两点绝对是学好矩阵分解的利器，怎么强调都不为过。有数据的地方，就有矩阵！有矩阵的地方，就有向量！有向量的地方，就有几何！

下面五章将展开讲解 Cholesky 分解、特征值分解和奇异值分解。本书最后三章会结合几何、数据、空间、应用等概念，再次升华矩阵分解！也就是说，有向量的地方，肯定有空间！