

24

Data Matrix Decomposition

数据分解

从几何、空间、优化、统计视角解读



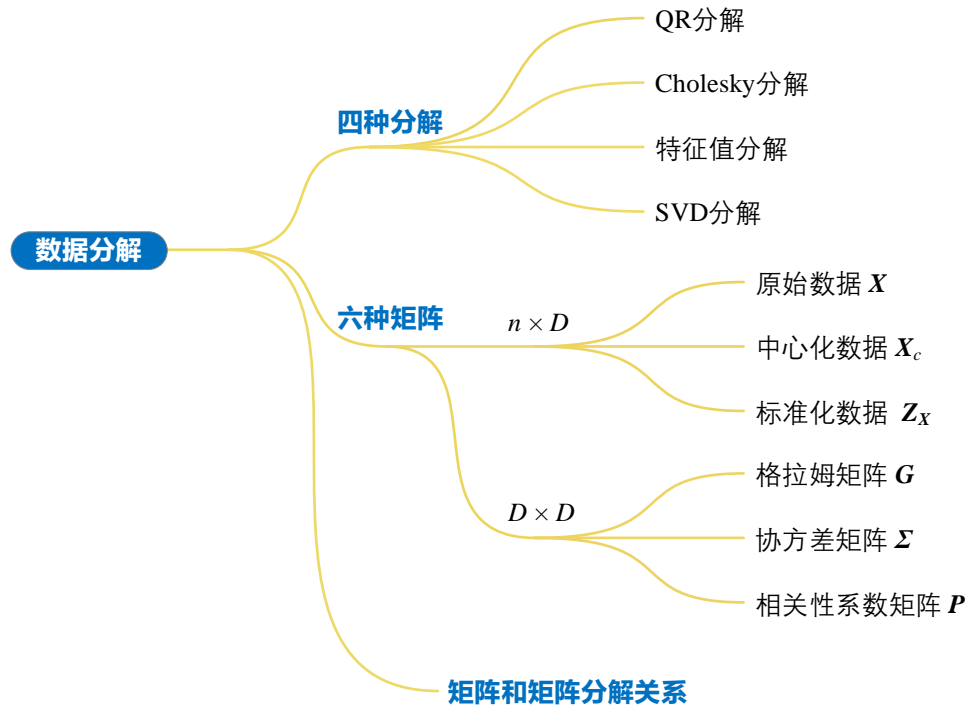
你不能教任何人任何东西，你只能帮助他在自己身上发现它。

You cannot teach a man anything; you can only help him discover it in himself.

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



- numpy.average() 计算平均值
- numpy.corrcoef() 计算数据的相关性系数
- numpy.cov() 计算协方差矩阵
- numpy.diag() 如果 A 为方阵，numpy.diag(A) 函数提取对角线元素，以向量形式输入结果；如果 a 为向量，numpy.diag(a) 函数将向量展开成方阵，方阵对角线元素为 a 向量元素
- numpy.linalg.cholesky() Cholesky 分解
- numpy.linalg.eig() 特征值分解
- numpy.linalg.inv() 矩阵求逆
- numpy.linalg.norm() 计算范数
- numpy.linalg.svd() 奇异值分解
- numpy.ones() 创建全 1 向量或矩阵
- numpy.sqrt() 计算平方根



24.1 为什么要分解矩阵？

QR 分解、Cholesky 分解、特征值分解、SVD 分解，这四种常用的分解的目的是什么？

它们分解的对象是什么？有何限制？

分解结果是什么？有何特殊性质？

矩阵分解之间有哪些区别和联系？

灵魂拷问来了——我们到底为什么需要分解矩阵？

大家可能会反问，前文学都学完了，现在才问是不是太晚了？

一点也不晚！矩阵分解是线性代数的核心中的核心，现在正是时候结合数据、几何、空间、优化、统计等视角总结这四种矩阵分解的最佳时机。

总结和比较

表 1 比较四种常用矩阵分解，请大家快速浏览这个表格，然后开始本章的学习。也请大家在完成本章内容学习后，回头仔细再看一遍表格内容。如果对任何矩阵分解细节感到生疏的话，请翻看本书前文对应内容。

表 1. 四种常用矩阵分解

矩阵分解	QR 分解	Cholesky 分解	特征值分解	SVD 分解
前提	任何实数矩阵都可以 QR 分解	正定矩阵才能 Cholesky 分解	可对角化矩阵才能进行特征值分解	任何实数矩阵都可以 SVD 分解
示意图				
公式	$A = QR$	$A = R^T R$ $A = LL^T$	$A = V \Lambda V^{-1}$ $A = V \Lambda V^T$ (A 为对称方阵时，其特征值分解又叫谱分解)	$A = USV^T$ (注意 V 的转置运算)
结果	Q 是正交矩阵 (完全型分解)，意味着 Q 是规范正交基 R 是上三角矩阵	L 为下三角方阵 R 为上三角方阵	Λ 为对角方阵，对角线元素为特征值 V 列向量为特征向量 如果 A 为对称方阵， V 为正交矩阵，即满足 $V^T V = V V^T = I$	U 为正交矩阵 (完全型分解)，它的列向量为左奇异向量 S 主对角线元素为奇异值 V 为正交矩阵 (完全型分解)，它的列向量为右奇异向量 U 和 V 都是规范正交基
几何视角	Q 代表旋转	写成 LDL^T 形式 (L 主对角线元素为 1)	V 代表旋转 Λ 代表缩放	U 代表旋转 S 代表缩放

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

		L 代表剪切 D 代表缩放		V 代表旋转
结果唯一?	A 列满秩, 且 R 的对角元素为正实数的情况下结果唯一	当限定 R 的对角元素为正时, 分解结果唯一	矩阵 V 不唯一 本书的特征向量都是单位向量, 特征向量一般差在正负符号上	矩阵 U 和 V 不唯一 本书左奇异向量、右奇异向量都是列向量
特殊类型	完全型 (Q 是正交矩阵) 经济型 (Q 是规范正交基, 但不是正交矩阵)	正定矩阵 埃尔米特矩阵 (不在本书讨论范围)	对称矩阵 正规矩阵 (不在本书讨论范围之内)	完全型 经济型 缩略型 截断型
向量空间	Q 的列向量为规范正交基, Q 的第一列向量 q_1 是 A 的第一列向量 a_1 的单位化 R 的列向量相当于坐标值	如果 $A = X^T X$ (即 Gram 矩阵) 正定, 对 A 进行 Cholesky 分解得到上三角矩阵 R , R 的列向量可以代表 X 列向量	如果 A 为对称方阵, V 为规范正交基 如果 $A = X^T X$ 且 X 列满秩, V 是 X 的行空间 $R(X)$	完全型 SVD 分解获得四个空间: 列空间 $C(X)$ 和左零空间 $\text{Null}(X^T)$, 行空间 $R(X)$ 和零空间 $\text{Null}(X)$ 完全型 SVD 分解相当于一次性完成两个特征值分解
优化视角			$\arg \max_v v^T A v$ subject to: $v^T v = 1$ 或 $\arg \max_{x \neq 0} \frac{x^T A x}{x^T x}$	$\arg \min_v \ A v\ $ subject to: $\ v\ = 1$ 或 $\arg \min_{x \neq 0} \frac{\ A x\ }{\ x\ }$
Numpy 函数	<code>numpy.linalg.qr()</code>	<code>numpy.linalg.cholesky()</code>	<code>numpy.linalg.eig()</code>	<code>numpy.linalg.svd()</code>
本章分解对象	原始数据矩阵 X	格拉姆矩阵 $G(X^T X)$ 协方差矩阵 Σ 相关性系数矩阵 P	格拉姆矩阵 $G(X^T X)$ 协方差矩阵 Σ 相关性系数矩阵 P	原始数据矩阵 X 中心化数据矩阵 X_c 标准化数据矩阵 Z_x
本系列丛书主要应用	解线性方程组 最小二乘回归 施密特正交化	蒙特卡罗模拟, 产生满足特定协方差矩阵要求的随机数 判断正定性	马尔科夫过程 主成分分析 瑞利商 矩阵范数	求解伪逆矩阵 矩阵范数 最小二乘回归 主成分分析 图像压缩

谱分解 \subset 特征值分解 \subset 奇异值分解

图 1 给出的文氏图为奇异值分解、特征值分解、谱分解之间的集合关系。

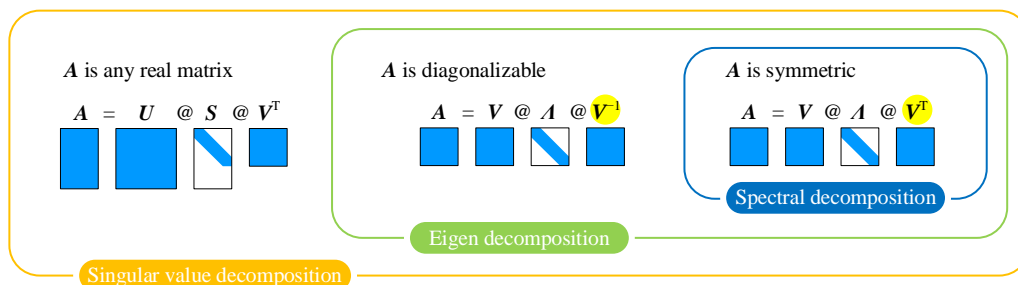


图 1. 特征值分解、奇异值分解之间的从属关系。

SVD 分解的对象是一切实数矩阵。特征值分解可以看做是特殊的 SVD 分解。特征值分解的对象是可对角化矩阵。如果 SVD 分解对象也是可对角化矩阵，其结果等价于特征值分解。注意，可对角化矩阵是特殊的方阵。

特别地，对称矩阵的特征值分解叫谱分解。格拉姆矩阵都是对称矩阵，因此格拉姆矩阵的特征值分解都是谱分解。谱分解得到的 V 是正交矩阵，正交矩阵是“天然”的规范正交基。从几何角度来看，正交矩阵的作用是“旋转”。

分解对象：数据矩阵，衍生矩阵

本章用的数据还是大家再熟悉不过的鸢尾花数据集。

快速回顾一下，如图 2 所示，鸢尾花数据矩阵 X 的每一列分别代表鸢尾花的不同特征——花萼长度（第 1 列，列向量 x_1 ）、花萼宽度（第 2 列，列向量 x_2 ）、花瓣长度（第 3 列，列向量 x_3 ）和花瓣宽度（第 4 列，列向量 x_4 ）。矩阵 X 的每一行代表一朵花的样本数据，每一行数据也是一个向量——行向量。注意，图 2 不考虑鸢尾花分类。

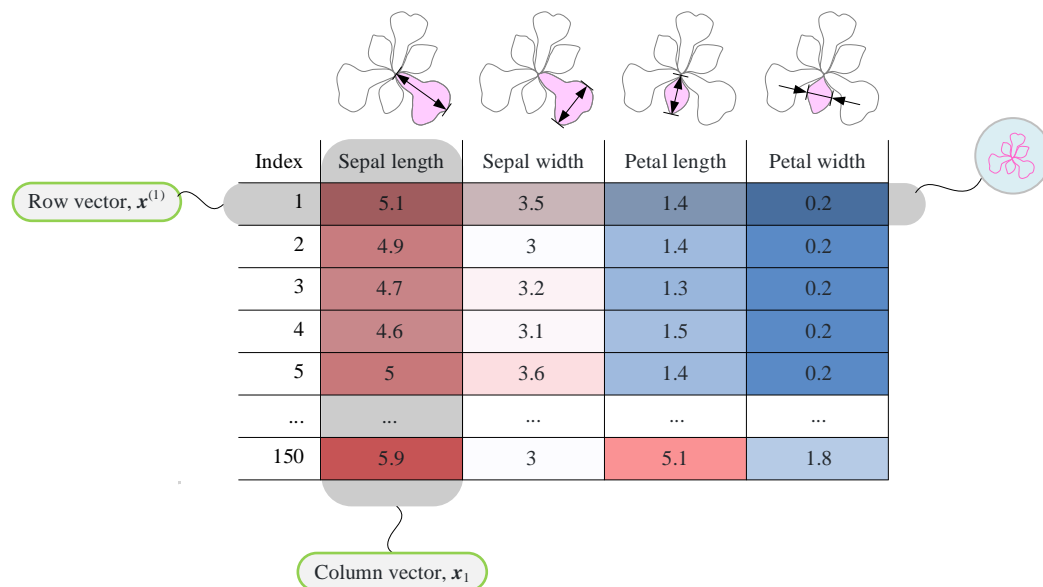


图 2. 鸢尾花数据集行、列含义

图 3 所示为本章矩阵分解对象，它们都衍生自鸢尾花数据矩阵 X 。 X 为细高长方形矩阵，形状为 $n \times D$ 。和本书第 10 章鸢尾花数据矩阵热图相比，图 3 中 X 的热图采用不同范围色谱。

格拉姆矩阵 G 来自数据矩阵 X ，两者关系为 $G = X^T X$ 。格拉姆矩阵 G 为对称矩阵。一般 G 为半正定，只有 X 满秩， G 才正定。

上一章提到，格拉姆矩阵 G 含有 X 列向量模、向量夹角两类重要信息。对于细高的长方形矩阵 X ，第二个格拉姆矩阵 XX^T 不常用。

而余弦相似度矩阵 C 仅有 X 列向量两两夹角信息。余弦相似度矩阵 C 也是对称矩阵。余弦相似度的取值 $[-1, 1]$ ，因此不同余弦相似度具有可比性。这一点类似统计中的相关性系数。对 X 列向量先进行单位化，再求其格拉姆矩阵，得到的就是 C 。

在统计视角下， X 的两个重要信息——质心 $E(X)$ (μ_X)、协方差矩阵 Σ_X (常简写为 Σ)。 $E(X)$ 对应数据质心位置， Σ_X 描述数据分布。注意，质心 $E(X)$ 、协方差矩阵 Σ_X 仅仅是描述数据矩阵 X 的统计工具而已，不代表 X 服从多元高斯分布 $N(\mu, \Sigma)$ 。

▲ 值得注意的是，本系列丛书定义 $E(X)$ 为行向量， $E(X)$ 的转置为列向量 μ_X 。

本章要用到两个和原始数据矩阵形状相同的矩阵——中心化数据矩阵 X_c 、标准化数据矩阵 Z_X 。 X 、 X_c 、 Z_X 的形状均为 $n \times D$ 。

X 每一列数据分别减去自己的均值便得到中心化数据 X_c ，即 $X_c = X - E(X)$ 。这个式子用到了广播原则。请大家回顾如何本书第 22 章有关如何用矩阵运算计算 X_c 。

几何视角，对于 X 来说，它的数据质心位于 μ_X ；而 X_c 的质心位于 0 。换个角度来看， X 的列向量起点位于原点；而 X_c 列向量的起点相当于移动到了质心，向量终点不动。

标准化数据 Z_X 实际上就是 X 的 z 分数。几何视角，从 X 到 Z_X 经过了平移、缩放两步操作。

▲ 注意，上一章创造了一个概念——标准差向量。标准差向量的模对应标准差大小，两个标准差向量的夹角余弦值对应相关性系数。

协方差矩阵 Σ_X 可以视作 X_c 的格拉姆矩阵。值得注意的是，计算 Σ_X 时使用了缩放系数 $1/n$ (总体) 或 $1/(n-1)$ (样本)。

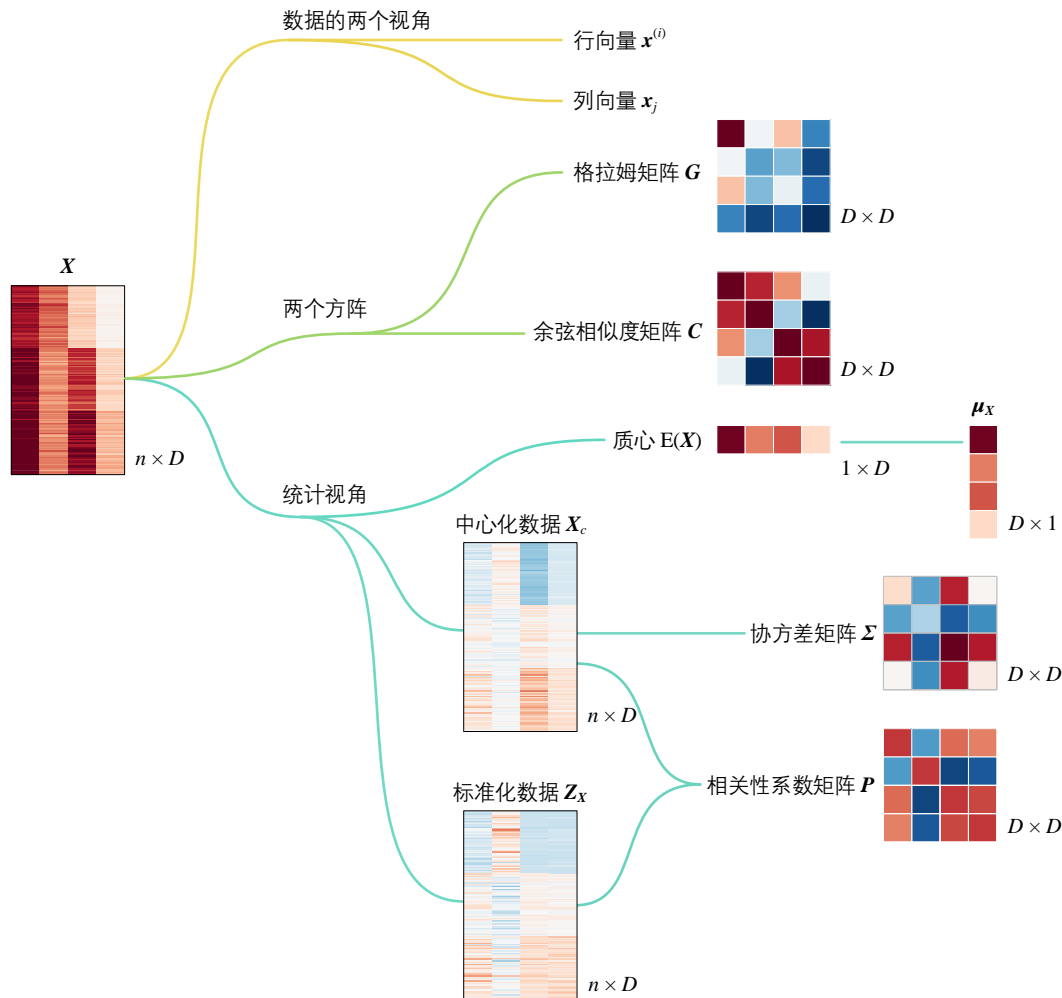
协方差矩阵 Σ 包含两类信息——标准差向量的模 (标准差)、标准差向量两两夹角 (相关性系数)。

相关性系数矩阵 P 仅仅含有标准差向量夹角 (相关性系数) 信息。相关性系数矩阵 P 类似余弦相似度矩阵 C 。

类似协方差矩阵 Σ ，计算相关性系数矩阵 P 也使用了缩放系数 $1/n$ (总体) 或 $1/(n-1)$ (样本)。相关性系数矩阵 P 就是标准化数据 Z_X 的协方差矩阵。

G 、 C 、 Σ 、 P 的形状均为 $D \times D$ 。

➡ 如果大家对这部分内容感到陌生，请回顾本书第 22 章。大家必须对矩阵分解的对象有足够的认识，才能开始本章后续内容学习。

图 3. X 衍生得到的几个矩阵

矩阵 + 矩阵分解

搭配六种不同矩阵 (X 、 X_c 、 Z_X 、 G 、 Σ 、 P) 和三种矩阵分解 (Cholesky 分解、特征值分解、SVD 分解)，会碰撞出什么？

表 2 给出了答案。本章后续内容将主要以表格中内容展开。

表 2. 矩阵和矩阵分解之间的关系

对象		Cholesky 分解	特征值分解	SVD 分解
$n \times D$	X	不适用	不适用	$X = U_X S_X V_X^T$
	$X_c = X - E(X)$	不适用	不适用	$X_c = U_c S_c V_c^T$
	$Z_X = (X - E(X)) S^{-1}$	不适用	不适用	$Z_X = U_Z S_Z V_Z^T$
	$S = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}}$	不适用	不适用	

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

$D \times D$	$G = X^T X$	正定矩阵为前提 $G = R_x^T R_x$	$G = V_x A_x V_x^T = V_x S_x^T S_x V_x^T$ S_x 来自于 X 的 SVD 分解	$G = V_x A_x V_x^T$
	样本: $\Sigma = \frac{(X - E(X))^T (X - E(X))}{n-1}$ $= \frac{X_c^T X_c}{n-1}$ 总体: $\Sigma = \frac{(X - E(X))^T (X - E(X))}{n}$ $= \frac{X_c^T X_c}{n}$	正定矩阵为前提 $\Sigma = R_c^T R_c$	样本: $\Sigma = V_c A_c V_c^T = V_c S_c^T S_c / (n-1) V_c^T$ 总体: $\Sigma = V_c A_c V_c^T = V_c S_c^T S_c / n V_c^T$ S_c 来自于 X_c 的 SVD 分解	$\Sigma = V_c A_c V_c^T$
	$P = S^{-1} \Sigma S^{-1}$ $S = \text{diag}(\text{diag}(\Sigma))^{\frac{1}{2}}$	正定矩阵为前提 $\Sigma = R_z^T R_z$	样本: $P = V_z A_z V_z^T = V_z S_z^T S_z / (n-1) V_z^T$ 总体: $P = V_z A_z V_z^T = V_z S_z^T S_z / n V_z^T$ S_z 来自于 Z_x 的 SVD 分解	$P = V_z A_z V_z^T$



Bk4_Ch24_01.py 中 Bk4_Ch24_01_A 部分计算得到图3所有矩阵，请读者根据前文所学自行绘制本章所有热图。

24.2 QR 分解：获得正交系

QR 分解不是本章重点，我们仅仅蜻蜓点水回顾一下。

如图4所示，对矩阵 X 进行缩略型 QR 分解，得到 Q 和 R 。 Q 和 X 形状相同，是正交矩阵的一部分，也就是说 Q 的列向量 $[q_1, q_2, q_3, q_4]$ 是规范正交基。 $[q_1, q_2, q_3, q_4]$ 相当于 $[x_1, x_2, x_3, x_4]$ 正交化的结果。

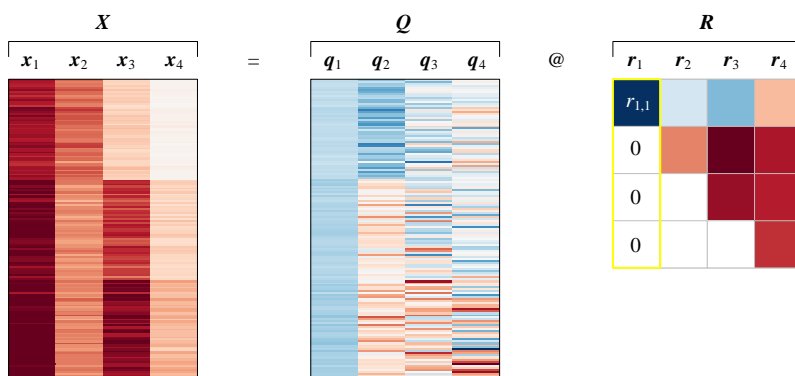


图 4. QR 分解热图

如图 5 所示，从空间角度来讲，如果 x_1, x_2, x_3, x_4 线性无关，则 $\text{span}(x_1, x_2, x_3, x_4) = \text{span}(q_1, q_2, q_3, q_4)$ 。

请大家特别关注如下关系：

$$x_1 = r_{1,1}q_1 \quad (1)$$

也就是说 x_1 和 q_1 平行。取决于 $r_{1,1}$ 正负， x_1 和 q_1 可以同向或反向。通过 QR 分解完成正交化相当于“顺藤”($x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4$) “摸瓜” ($q_1 \rightarrow q_2 \rightarrow q_3 \rightarrow q_4$)。 $(r_{1,1}, 0, 0, 0)$ 是 x_1 在基底 $[q_1, q_2, q_3, q_4]$ 的坐标。此外请大家注意，QR 分解和格拉姆-施密特正交化 (Gram-Schmidt process) 之间联系。

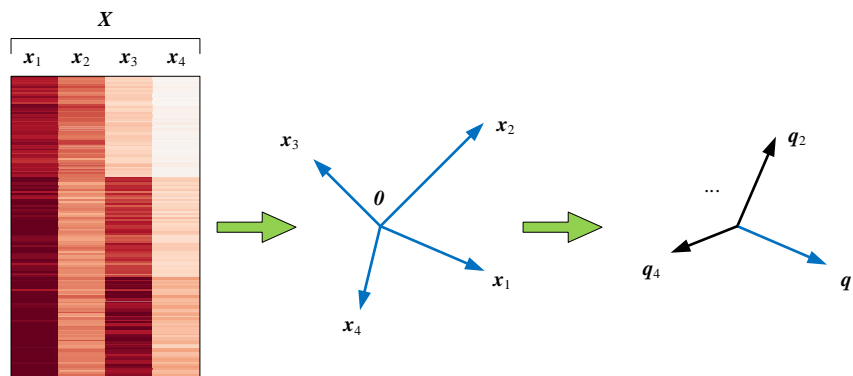
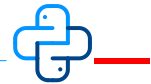


图 5. $[q_1, q_2, q_3, q_4]$ 是规范正交基



Bk4_Ch24_01.py 中 Bk4_Ch24_01_B 部分完成矩阵 X 的 QR 分解。

24.3 Cholesky 分解：找到列向量的坐标

格拉姆矩阵

数据矩阵 X 的每一列可以看做一个向量，而 Cholesky 分解能够找到它们的坐标。

▲ 注意，这里存在一个前提—— X 列满秩。只有这样 X 的格拉姆矩阵 G 才正定，才能进行 Cholesky 分解。

假设 G 正定，对 G 进行 Cholesky 分解：

$$G = R^T R \quad (2)$$

其中， R 为上三角矩阵。(2) 中的 R 不同于上一节 QR 分解的 R 。

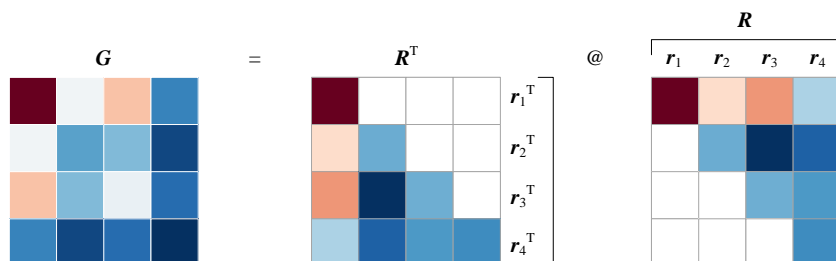


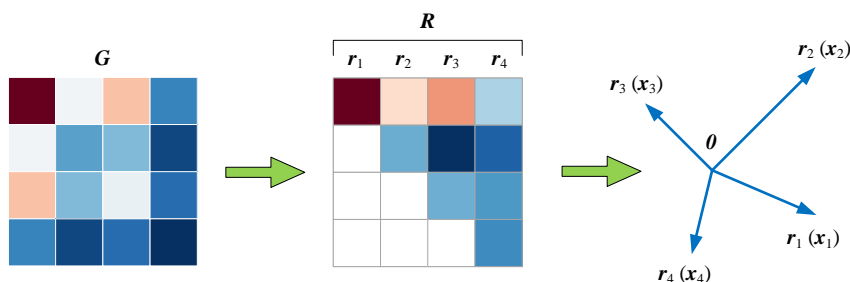
图 6. 对格拉姆矩阵 G 进行 Cholesky 分解矩阵运算热图

如图 6 所示，将 R 写成 $[r_1, r_2, \dots, r_D]$ ，(2) 可以写成向量标量积形式，并建立它们和 $[x_1, x_2, \dots, x_D]$ 的联系：

$$\begin{aligned}
 G = R^T R &= \begin{bmatrix} \langle r_1, r_1 \rangle & \langle r_1, r_2 \rangle & \cdots & \langle r_1, r_D \rangle \\ \langle r_2, r_1 \rangle & \langle r_2, r_2 \rangle & \cdots & \langle r_2, r_D \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle r_D, r_1 \rangle & \langle r_D, r_2 \rangle & \cdots & \langle r_D, r_D \rangle \end{bmatrix} = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_D \rangle \\ \langle x_2, x_1 \rangle & \langle x_2, x_2 \rangle & \cdots & \langle x_2, x_D \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle x_D, x_1 \rangle & \langle x_D, x_2 \rangle & \cdots & \langle x_D, x_D \rangle \end{bmatrix} \\
 &= \begin{bmatrix} \|r_1\| \|r_1\| \cos \theta_{1,1} & \|r_1\| \|r_2\| \cos \theta_{2,1} & \cdots & \|r_1\| \|r_D\| \cos \theta_{1,D} \\ \|r_2\| \|r_1\| \cos \theta_{1,2} & \|r_2\| \|r_2\| \cos \theta_{2,2} & \cdots & \|r_2\| \|r_D\| \cos \theta_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \|r_D\| \|r_1\| \cos \theta_{1,D} & \|r_D\| \|r_2\| \cos \theta_{2,D} & \cdots & \|r_D\| \|r_D\| \cos \theta_{D,D} \end{bmatrix} \\
 &= \begin{bmatrix} \|x_1\| \|x_1\| \cos \theta_{1,1} & \|x_1\| \|x_2\| \cos \theta_{2,1} & \cdots & \|x_1\| \|x_D\| \cos \theta_{1,D} \\ \|x_2\| \|x_1\| \cos \theta_{1,2} & \|x_2\| \|x_2\| \cos \theta_{2,2} & \cdots & \|x_2\| \|x_D\| \cos \theta_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \|x_D\| \|x_1\| \cos \theta_{1,D} & \|x_D\| \|x_2\| \cos \theta_{2,D} & \cdots & \|x_D\| \|x_D\| \cos \theta_{D,D} \end{bmatrix} \quad (3)
 \end{aligned}$$

$[r_1, r_2, \dots, r_D]$ 的每个列向量的模分别等于 $[x_1, x_2, \dots, x_D]$ 列向量的模； $[r_1, r_2, \dots, r_D]$ 中两两向量夹角等于 $[x_1, x_2, \dots, x_D]$ 中对应列向量夹角。

换个角度来看， X 的形状为 $n \times D$ ，比如 150×4 。 X 的 4 个列向量为 150 维，“装下”这些列向量我们自然先考虑 \mathbb{R}^{150} 空间。而 R 的形状为 4×4 ，用 \mathbb{R}^4 空间装下 R 列向量刚刚好。“刚刚好”是因为 R 满秩。也就是说，我们用 \mathbb{R}^4 空间中的 $[r_1, r_2, r_3, r_4]$ 来“代表” \mathbb{R}^{150} 空间中的 $[x_1, x_2, x_3, x_4]$ 。显然， R 远比 X “小巧”的多。


图 7. $[x_1, x_2, x_3, x_4]$ 和 $[r_1, r_2, r_3, r_4]$ “等价”

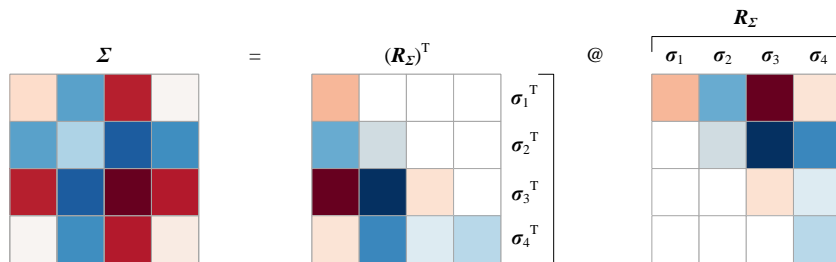
协方差矩阵

类似地，对协方差矩阵 Σ 进行 Cholesky 分解，具体如图 8 所示。将 R_Σ 写成“标准差向量” $[\sigma_1, \sigma_2, \dots, \sigma_D]$ ，整理得到：

$$\Sigma = R_\Sigma^T R_\Sigma = \begin{bmatrix} \langle \sigma_1, \sigma_1 \rangle & \langle \sigma_1, \sigma_2 \rangle & \cdots & \langle \sigma_1, \sigma_D \rangle \\ \langle \sigma_2, \sigma_1 \rangle & \langle \sigma_2, \sigma_2 \rangle & \cdots & \langle \sigma_2, \sigma_D \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \sigma_D, \sigma_1 \rangle & \langle \sigma_D, \sigma_2 \rangle & \cdots & \langle \sigma_D, \sigma_D \rangle \end{bmatrix} = \begin{bmatrix} \text{cov}(X_1, X_1) & \text{cov}(X_1, X_2) & \cdots & \text{cov}(X_1, X_D) \\ \text{cov}(X_2, X_1) & \text{cov}(X_2, X_2) & \cdots & \text{cov}(X_2, X_D) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(X_D, X_1) & \text{cov}(X_D, X_2) & \cdots & \text{cov}(X_D, X_D) \end{bmatrix} \quad (4)$$

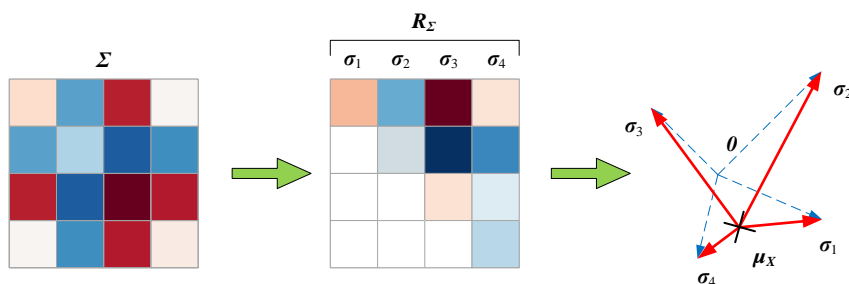
当然，我们也可以对线性相关系数矩阵 P 进行 Cholesky 分解。

 R_Σ 将会用在蒙特卡洛模拟中，用来生成满足协方差矩阵 Σ 要求的随机数组，这是本系列丛书《概率统计》要讨论的内容。

图 8. 对协方差矩阵 Σ 进行 Cholesky 分解矩阵运算热图

向量 $\sigma_1, \sigma_2, \dots, \sigma_D$ 的模分别对应 $x_1 (X_1), x_2 (X_2), \dots, x_D (X_D)$ 的标准差，向量 $\sigma_1, \sigma_2, \dots, \sigma_D$ 两两夹角余弦值对应 $x_1 (X_1), x_2 (X_2), \dots, x_D (X_D)$ 的两两线性相关系数。也就是说，协方差矩阵 Σ 集成了标准差和线性相关系数这两类信息。

如图 9 所示， $[\sigma_1, \sigma_2, \dots, \sigma_D]$ 相当于以数据 X 质心为中心一组非正交基。数据 X 的很多统计学运算和分析都是依托这个空间完成的。

图 9. $[\sigma_1, \sigma_2, \dots, \sigma_D]$ 相当于以 X 质心为中心张成一个空间

Bk4_Ch24_01.py 中 Bk4_Ch24_01_c 部分完成对格拉姆矩阵 G 和协方差矩阵 Σ 的 Cholesky 分解。

24.4 特征值分解：获得行空间和零空间

本节要进行三个特征值分解，为了区分，我们在分解结果加了下角标。

格拉姆矩阵

图 10 所示为格拉姆矩阵 $G = X^T X$ 进行特征值分解。因为 G 为对称矩阵，所以 V_X 为正交矩阵，即满足 $V_X^{-1} = V_X^T$ 。从而， G 的特征值分解可以写成 $G = V_X \Lambda_X V_X^T$ 。

根据上一章内容， V_X 的列向量 $[v_{X_1}, v_{X_2}, \dots, v_{X_D}]$ 是一组规范正交基。 $[v_{X_1}, v_{X_2}, \dots, v_{X_D}]$ 张成 \mathbb{R}^D 空间，它是矩阵 X 的行空间和零空间的合体。零空间的维数取决于 G 的秩。

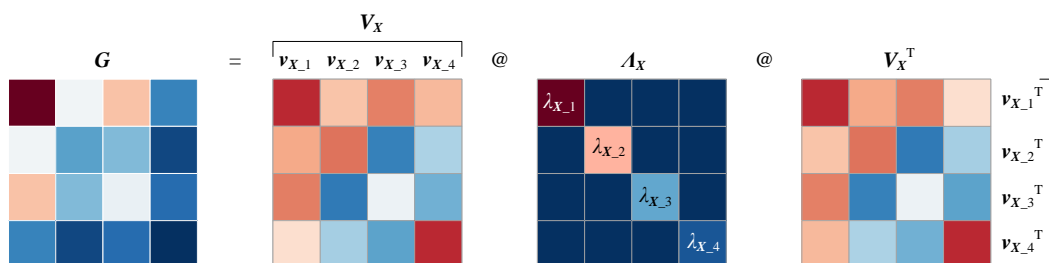


图 10. 对格拉姆矩阵进行特征值分解

如图 11 所示，从 X 到 V_X 相当于对 X 行向量正交化。根据前文所学，大家思考以下几个问题。

X 投影到 v_{X_1} 的结果怎么计算？ X 投影到 V_X 的结果又怎么计算？投影结果有怎样性质？

值得注意的是，本章矩阵 X 为鸢尾花数据，每一列数据单位都是厘米 (cm)。格拉姆矩阵 G 中数值的单位为平方厘米 cm^2 。 V_X 中每一列都是单位向量，仅仅表达方向，不含有单位。而特征值 λ_x 的单位为平方厘米 cm^2 。从几何角度来看，特征值含有椭圆 (椭球) 的大小形状信息，而 V 仅提供空间旋转操作。

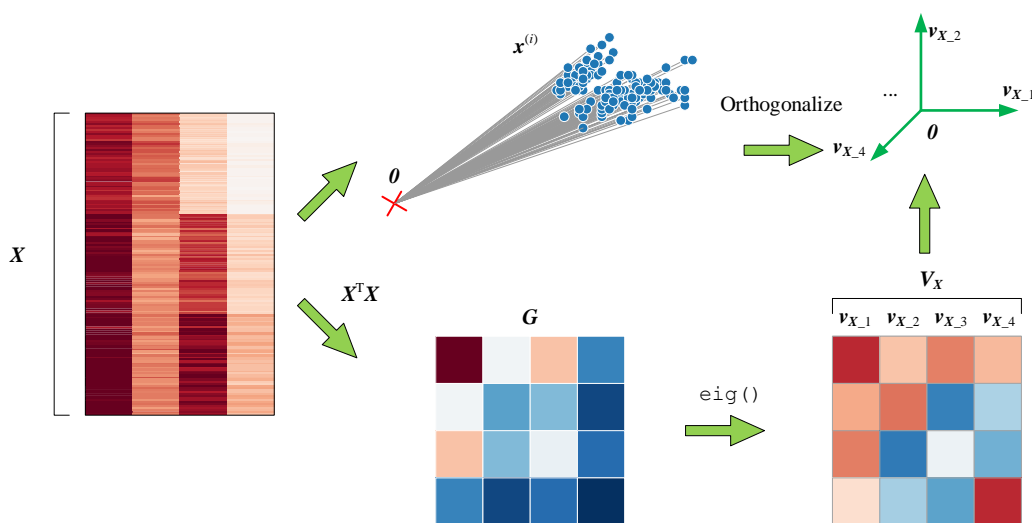


图 11. 特征值分解 G 获得规范正交基

优化视角

本书第 18 章讲过，获得规范正交基 $[v_{X_1}, v_{X_2}, \dots, v_{X_D}]$ 有着特定的优化目标。下面，我们简要回顾一下。

矩阵 X 在 v 方向投影得到 y ：

$$Xv = y \quad (5)$$

而 $v^T G v$ 可以写成：

$$v^T G v = v^T X^T X v = (Xv)^T Xv = y^T y = \|y\|_2^2 \quad (6)$$

这就是特征值分解格拉姆矩阵对应的优化问题——找到一个单位向量 v ，使得 X 在 v 上投影结果 y 的模最大。这个 v 就是 v_{X_1} ，对应 y 的模最大值为 $\sqrt{\lambda_{X_1}}$ 。

而 y 的模的平方 $\|y\|_2^2$ 就是 y 中所有坐标点距离原点的欧氏距离平方。

解决这个优化问题采用的方法可以是本书第 14 章讲的瑞利商，也可以是第 18 章讲的拉格朗日乘子法。两者在本质上是是一致的。

有了 \mathbf{v}_{X_1} ，寻找 \mathbf{v}_{X_2} 时，首先让 \mathbf{v}_{X_2} 垂直 \mathbf{v}_{X_1} (约束条件)，且 \mathbf{X} 在 \mathbf{v}_{X_2} 上投影结果 \mathbf{y} 的模最大。以此类推得到所有特征向量、特征值。

特征值

前文介绍过，特征值分解得到的特征值之和，等于原矩阵对角线元素之和，即：

$$\lambda_{X_1} + \lambda_{X_2} + \lambda_{X_3} + \lambda_{X_4} = \text{sum}(\text{diag}(\mathbf{G})) = \|\mathbf{x}_1\|_2^2 + \|\mathbf{x}_2\|_2^2 + \|\mathbf{x}_3\|_2^2 + \|\mathbf{x}_4\|_2^2 \quad (7)$$

协方差矩阵

第二个例子是对协方差矩阵 Σ 进行特征值分解，图 12 所示为对应热图。下角标用“c”的原因是对协方差矩阵特征值分解结果和中心化 (去均值) 矩阵 \mathbf{X}_c 直接相关。

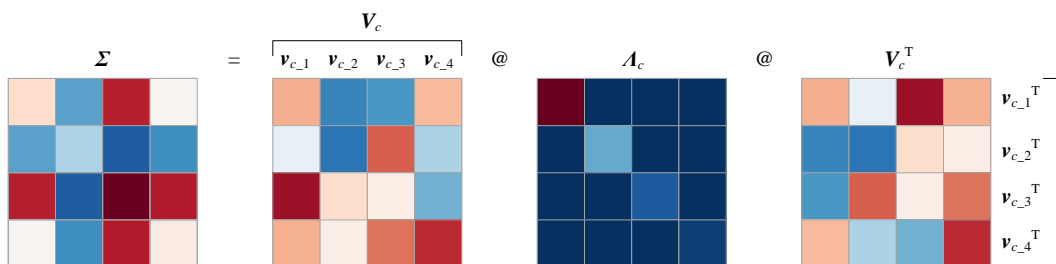


图 12. 对协方差矩阵进行特征值分解

前文提到过， Σ 囊括标准差向量 $[\sigma_1, \sigma_2, \sigma_3, \sigma_4]$ 所有信息——模 (标准差) 和夹角余弦值 (线性相关系数)。对 Σ 特征值分解得到的特征向量矩阵 $[\mathbf{v}_{c,1}, \mathbf{v}_{c,2}, \mathbf{v}_{c,3}, \mathbf{v}_{c,4}]$ 也是一组规范正交基，它显然不同于 $[\mathbf{v}_{X_1}, \mathbf{v}_{X_2}, \dots, \mathbf{v}_{X_D}]$ 。

大家思考一个问题， \mathbf{X} 和 \mathbf{X}_c 分别向 $[\mathbf{v}_{c,1}, \mathbf{v}_{c,2}, \mathbf{v}_{c,3}, \mathbf{v}_{c,4}]$ 和 $[\mathbf{v}_{X_1}, \mathbf{v}_{X_2}, \dots, \mathbf{v}_{X_D}]$ 投影产生的 4 种结果有怎样差别？

优化视角

采用和本节前文一样的优化角度分析协方差矩阵的特征值分解。

中心化数据矩阵 \mathbf{X}_c 向 \mathbf{v} 投影得到 \mathbf{y}_c ：

$$\mathbf{X}_c \mathbf{v} = \mathbf{y}_c \quad (8)$$

而 $\mathbf{v}^T \Sigma \mathbf{v}$ 可以写成：

$$\mathbf{v}^T (n-1) \Sigma \mathbf{v} = \mathbf{v}^T \mathbf{X}_c^T \mathbf{X}_c \mathbf{v} = (\mathbf{X}_c \mathbf{v})^T \mathbf{X}_c \mathbf{v} = \mathbf{y}_c^T \mathbf{y}_c = \|\mathbf{y}_c\|_2^2 = (n-1) \text{var}(\mathbf{y}_c) \quad (9)$$

上式告诉我们，对协方差矩阵特征值分解，就是要找到一个单位向量 \mathbf{v} ，使得中心化数据 \mathbf{X}_c 在 \mathbf{v} 上投影结果 \mathbf{y}_c 的方差最大。我们要找的这个 \mathbf{v} 就是图 12 中的 $\mathbf{v}_{c,1}$ ，对应的特征值为 $\lambda_{c,1}$ 。

再次注意单位问题，对于鸢尾花数据，协方差矩阵中的数值单位都是平方厘米 cm^2 。其特征值 λ_c 的单位也是平方厘米 cm^2 ，而 \mathbf{v}_c 是无单位的。

➡ 大家可能会问，(9) 是如何把协方差矩阵和 \mathbf{y} 的方差联系起来的？这是我们下一章要探讨的内容。

Σ 的特征值之和，等于 \mathbf{X} 的每列数据方差之和，即：

$$\lambda_{x,1} + \lambda_{x,2} + \lambda_{x,3} + \lambda_{x,4} = \text{diag}(\Sigma) = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + \sigma_4^2 \quad (10)$$

显然 $\lambda_{x,1}$ 在 (10) 中占比最大。也就是说，对 Σ 特征值分解得到第一特征向量 $\mathbf{v}_{c,1}$ ，相较其他所有可能的单位向量，解释了 Σ 中最多的方差成分。

➡ 每个特征值占特征值总和的比例是主成分分析中重要的一项分析指标，这是本系列丛书《数据科学》一册要介绍的内容。

相关性系数矩阵

本节的第三个例子是对相关性系数矩阵 \mathbf{P} 特征值分解，图 13 所示为对应热图。相关性系数矩阵 \mathbf{P} 可以视作 \mathbf{Z}_x (\mathbf{X} 的 z 分数矩阵) 的协方差矩阵。

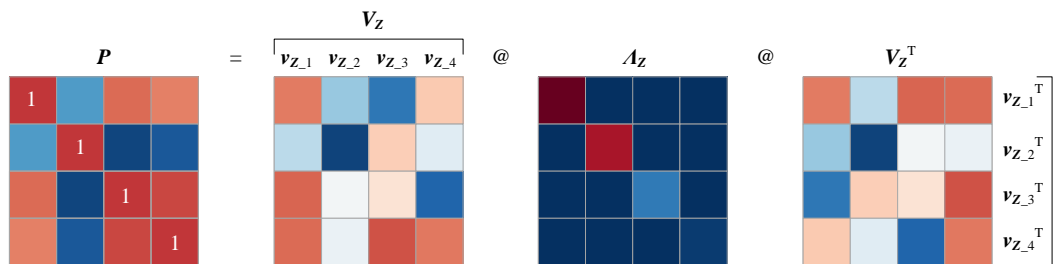


图 13. 对相关性系数矩阵进行特征值分解

矩阵 \mathbf{Z}_x 的特点是，每列均值都是 0。由于 \mathbf{Z}_x 已经标准化，每列的均方差为 1。矩阵 \mathbf{Z}_x 的列向量可以看成是一排单位向量 $\left[\frac{\sigma_1}{\|\sigma_1\|}, \frac{\sigma_2}{\|\sigma_2\|}, \frac{\sigma_2}{\|\sigma_2\|}, \frac{\sigma_2}{\|\sigma_2\|} \right]$ 。

从相关性系数矩阵 \mathbf{P} 对角线元素也可以看出来， \mathbf{Z}_x 每个特征贡献的方差均为 1。

⚠ 注意，数据矩阵 \mathbf{Z}_x 和相关性系数矩阵 \mathbf{P} 都已经“去单位化”。比如， \mathbf{P} 中对角线上的 1 没有单位；因为数据标准化的过程，单位已经消去。

对 P 的特征向量矩阵 $[vz_1, vz_2, vz_3, vz_4]$ 也是一组规范正交基。一般情况, $[vz_1, vz_2, vz_3, vz_4]$ 不同于 $[vc_1, vc_2, vc_3, vc_4]$ 。

利用对相关系数矩阵特征值分解进行主成分分析也是常见技术路线。这种技术路线可以解决 X 中某些特征的方差异常 (过大或过小) 的问题。



Bk4_Ch24_01.py 中 Bk4_Ch24_01_D 部分完成本节介绍的三个特征值分解。

24.5 SVD 分解：获得四个空间

SVD 分解可谓矩阵分解之集大成者，本书前文花了很多笔墨从各个角度探讨 SVD 分解。本节对比鸢尾花原始数据矩阵 X 、中心化矩阵 X_c 、标准化矩阵 Z_X 等三个矩阵 SVD 分解。

原始数据矩阵

图 14 所示为矩阵 X 进行 SVD 分解矩阵运算热图。图中的正交矩阵 V_X 实际上和图 10 中的 V_X 等价，某些向量的正负号可能存在反号的情况。图 14 中矩阵 U_X 也可以通过对 XX^T 特征值分解得到。

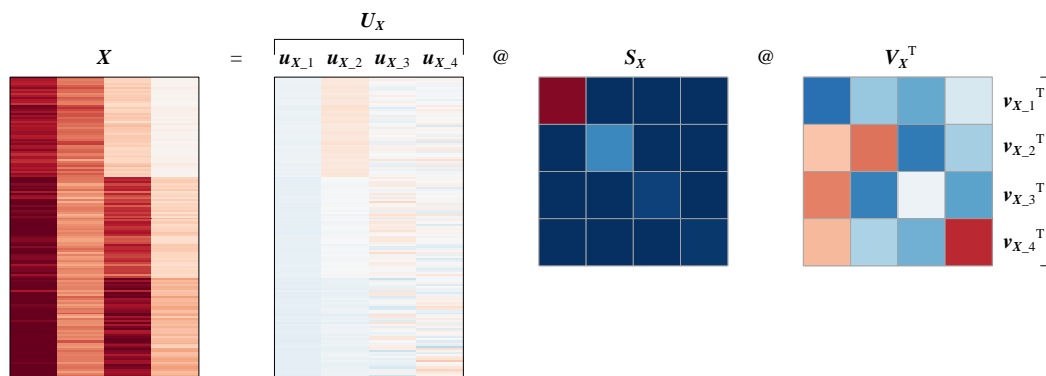


图 14. 对矩阵 X 进行 SVD 分解

前文提到多次，SVD 分解的结果包含了两个特征值分解结果。此外，SVD 分解不丢失原始数据 X 的任何信息，截断型除外。某种程度上说，SVD 分解包含了特征值分解，比特征值分解更“高阶”。

另外，请大家注意图 14 中奇异值和图 10 中特征值之间的关系：

$$S_X^2 = A_X \quad (11)$$

中心化数据

图 15 所示为中心化数据矩阵 X_c 进行 SVD 分解矩阵运算热图。图 15 中的正交矩阵 V_c 和图 12 中的 V_c 等价，两者若干位置列向量也可能存在符号相反情况。

有些读者可能会问，既然 V_c 也是规范正交基，那么将原始数据 X 在 V_c 上投影结果的质心在哪？投影结果的协方差矩阵又如何？下一章会给大家一些理论基础，本系列丛书《概率统计》一册会专门回答这个问题。

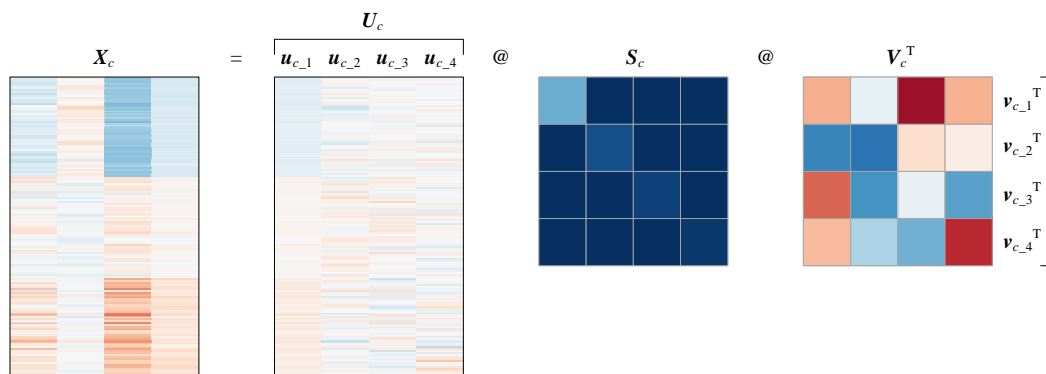


图 15. 对矩阵 X_c 进行 SVD 分解

标准化数据

图 16 所示为标准化数据矩阵 Z_X 进行 SVD 分解矩阵运算热图。图中的 V_Z 和图 13 中的 V_Z 等价，两者某些列向量也可能存在符号相反情况。也请大家思考，原始数据 X 在 V_Z 上投影会有怎样的结果？

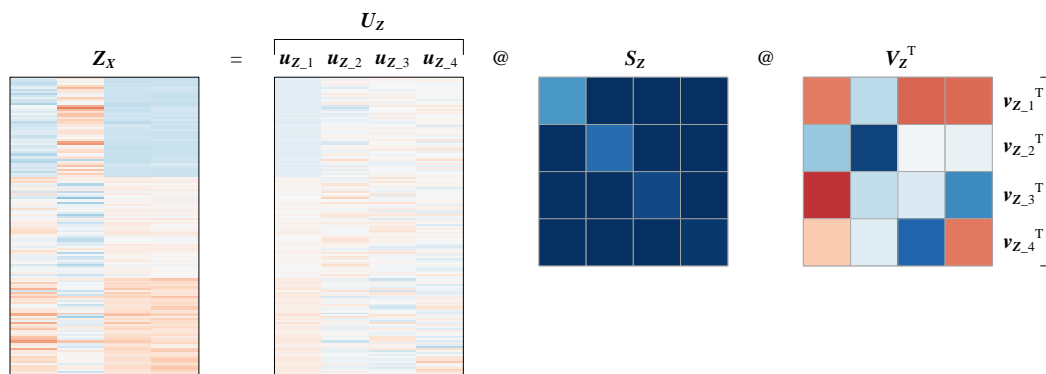
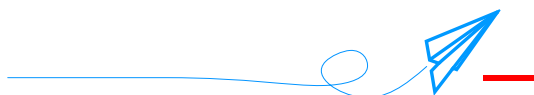


图 16. 对矩阵 Z_X 进行 SVD 分解



Bk4_Ch24_01.py 中 Bk4_Ch24_01_E 部分完成本节三个 SVD 分解运算。



本章最后用一幅图总结本章和上一章内容。

图 17 这幅图是本书中非常重要的几幅图之一，这幅图总结了整本书中和数据矩阵 X 有关的向量、矩阵、矩阵分解、空间等概念。

这幅图的数据分为两个部分：第一部分以 X 为核心，向量以 0 为起点；第二部分是统计视角，以去均值数据 X_c 为核心，向量以质心为起点。

下面，我们聊一下图 17 中关键细节。

X 为细高型矩阵，形状为 $n \times D$ ，样本数 n 一般远大于特征数 D 。对 X 进行 SVD 分解可以得到四个空间。

行空间 $R(X)$ “刚刚好”装下 X 的行向量。而 \mathbb{R}^D 装下 X 行向量后则可能略有富余，多余的部分就是零空间 $\text{Null}(X)$ 。零空间维数大于 0 的前提是 X 非满秩。

同理，列空间 $C(X)$ 正好装下 X 的列向量，没有富余。而 \mathbb{R}^n 装 X 的列向量则“富富有余”，“有余”的部分就是左零空间 $\text{Null}(X^T)$ 。

格拉姆矩阵 G 含有 X 列向量模、向量夹角两类重要信息。余弦相似度矩阵 C 仅仅含有向量夹角信息。对格拉姆矩阵 G 进行特征值分解只能获得两个空间。

对格拉姆矩阵 G 进行 Cholesky 分解得到上三角矩阵 R 可以“代表” X 列向量坐标。

▲ 反复强调，只有正定矩阵才能进行 Cholesky 分解。

在统计视角下， X 有两个重要信息——质心、协方差矩阵。质心确定数据中心位置，协方差矩阵描述数据分布。协方差矩阵 Σ 同样含有“标准差向量”的模（标准差大小）、向量夹角（余弦值为相关性系数）两类重要信息。相关性系数矩阵 P 仅仅含有向量夹角（相关性系数）信息。

▲ 值得格外注意的是，质心和协方差是多元高斯分布的两个参数，因此需要大家注意协方差矩阵和椭圆的联系。对这部分内容生疏的读者，请参考本书第 14 章。

X_c 是中心化数据矩阵，即每一列数据都去均值。 Z_x 是标准化数据矩阵，即 X 的 z 分数。在几何视角下， X 到 X_c 相当于质心“平移”， X 到标准化数据 Z_x 相当于“平移 + 缩放”。

协方差矩阵 Σ 相当于 X_c 的格拉姆矩阵。相关性系数矩阵 P 相当于 Z_x 的格拉姆矩阵。此外，注意样本数据缩放系数 $(n-1)$ 。

X_c 进行 SVD 分解也得到四个空间。这四个空间因 X_c 而生，一般情况不同于 X 的四个空间。

此外，请大家格外注意不同矩阵的单位！以鸢尾花数据为例， X 的每一列数据单位恰好都是 cm， X_c 的单位也都是 cm，而 Z_x 没有单位（或者说，单位是标准差）； G 的单位是 cm^2 ， Σ 的单位也是 cm^2 ， P 和 C 没有单位。

但是多数时候数据矩阵列向量特征比较丰富，比如高度、质量、时间、温度、密度、百分比、股价、收益率、GDP 等等。它们的数值单位不同、取值范围不同、均方差不同，为了保证可比性，我们需要标准化处理原始数据。

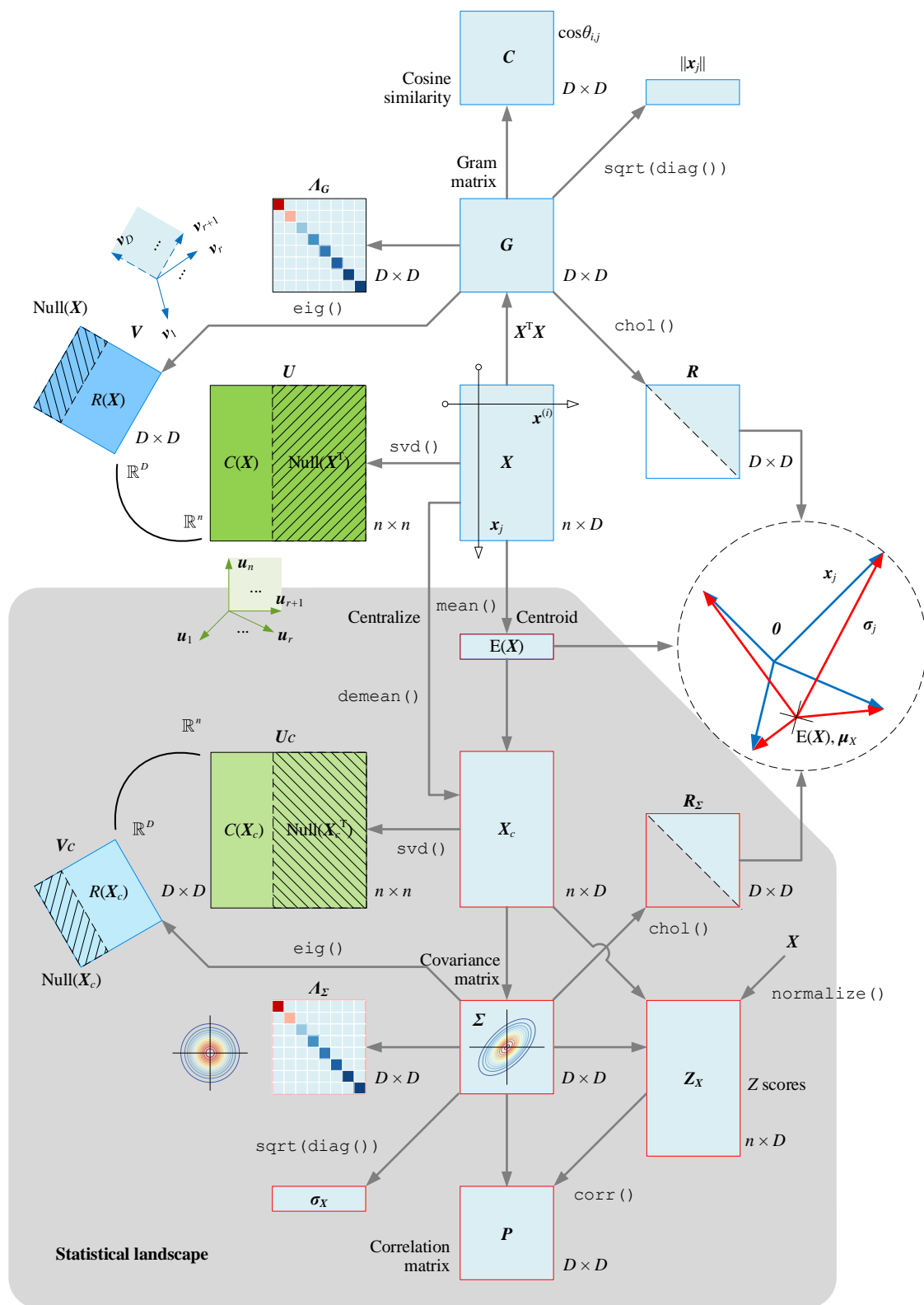


图 17. 总结本章内容的一幅图